DOCUMENT RESUME

ED 185 089 TM 800 088

AUTHOR Medley, Donald M.: Hill, Russell A.
TITLE Measurement Properties of Observation Schedule and Record, Form 4V.
PUB DATE [73]
NOTE 132p.: Some tables may be marginally legible.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Analysis of Variance: Behavior Rating Scales: *Classroom Observation Techniques: Classroom Research: Factor Structure: Interaction Process Analysis: Lesson Observation Criteria: *Scoring: Secondary Education: *Speech Communication: Student Teacher Relationship: *Teacher Behavior: Teacher Effectiveness: Test Reliability
IDENTIFIERS Flanders System of Interaction Analysis: National Teacher Examinations: *Observation Schedule and Record

ABSTRACT

The rationale of the Observation Schedule and Record (OScAR) is described, and research is reported to show the extent to which the form fulfills its purposes. In order to provide objective, reliable, and valid information about the teaching process, this classroom observation schedule is used to record the frequency of occurrence of 42 teacher behaviors only. The scores for OScAR are based on 41 contrast keys showing the relative frequencies among sets of the 42 observed behaviors. Eighteen keys are based on contrasts among statements and non-substantive interchanges, 3 keys on entries to substantive interchanges, 5 keys on exits from substantive interchanges, and 15 keys on interaction entries and exits. Research on the schedule is based on a study of 70 teacher interns in secondary schools, with four observations of each teacher. Simultaneously other observers recorded the same situation using the Flanders' Interaction Analysis Technique. Analysis of variance was used to estimate score reliabilities and factor analysis of both rating forms together was used for score interpretation. Scores were also related to subscores on the National Teacher Examination. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED185089

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Measurement Properties of Observation Schedule and Record, Form 4V

Donald M. Medley

Educational Testing Service, Princeton, New Jersey[1]

Russell A. Hill

Temple University, Philadelphia, Pennsylvania[2]

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Donald M. Medley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 088

Running Head: Measurement Properties of Observation Schedule and Record

[1]Now at the University of Virginia, Charlottesville, Virginia

[2]Now at Research for Better Schools, Philadelphia, Pennsylvania

TABLE OF CONTENTS

	Page
OScAR 4V	9
DESCRIPTION OF THE CATEGORIES	11
SCORING PROCEDURES FOR OScAR 4V	15
A NOTE ON ORTHOGONAL CONTRASTS	19
THE SET OF CONTRAST KEYS	24
FLANDERS' INTERACTION ANALYSIS TECHNIQUE (FIAT)	30
PRELIMINARY ANALYSIS OF DATA	33
COLLECTION OF DATA	33
PRELIMINARY ANALYSIS OF OScAR DATA	35
PRELIMINARY ANALYSIS OF FIAT DATA	40
FACTOR ANALYSIS	42
DISCUSSION	49
ANALYSIS OF VARIANCE OF FACTOR SCORES	55
CHANGES IN BEHAVIOR	55
SUBJECT-MATTER DIFFERENCES	59
RELIABILITY OF CHANGE SCORES	61
COGNITIVE FACTORS IN TEACHER BEHAVIOR	63
THE EXAMINATION	63
ANALYSIS OF INTERNAL STRUCTURE OF THE EXAMINATION	64
REGRESSION ON BEHAVIOR DIMENSIONS	66
DISCUSSION	68
SUMMARY AND CONCLUSIONS	70
REFERENCES	73

Measurement Properties of Observation Schedule and Record, Form 4V

One of the chief obstacles to the improvement of teaching is a lack of a precise, objective language in terms of which the phenomenon can be analyzed and described. The evolution of teaching from a little-understood, difficult-to-communicate art to a systematic, dynamic science depends on the development of such a language. This language of teacher behavior (Medley, 1971) must be based on terms which have overt, observable referents and which can be quantified for analysis if it is to become the basis of a science of teaching. And only through the development of such a science can it ever become possible effectively to apply the scientific method to the improvement of public education. Only when this powerful tool becomes available can there be significant and rapid improvement in the process of teaching--a process which is and will remain for a long time at the very heart of public education.

The term teaching is used here in a somewhat different sense than is usual today. Teaching is here defined as a process by which one human being (called the teacher) guides the learning of a heterogeneous group of other human beings (called pupils) in such a manner as to develop, not only those changes in behavior whose production constitute the immediate objectives of instruction, but also a set of learning skills, habits, and attitudes, the possession of which will gradually render the teacher

unnecessary. The aim of teaching may be said to be the transforming of pupils into educated men and women, an educated person being conceived of as one who possesses knowledge, the ability to learn, the habit of learning, and a thirst for learning.

Teaching is, then, something done by a human teacher. The functions performed by teaching machines or computers are best referred to as tutoring or instruction. These terms are perhaps most useful when distinguished as follows: tutoring is a process by which one person assists another to learn through mutual interaction--that is, through a process in which either tutor or learner may initiate interaction. Instruction is a process by which content is communicated unilaterally by an instructor in a manner designed to facilitate learning on the part of the pupil. There is no interaction in instruction.

Computer assisted instruction programs tutor rather than teach. Lecturers, educational films, closed-circuit television courses, and most programmed learning materials instruct rather than teach. Unlike tutoring, which is an individual process, or teaching, which is a group process, instruction can go on with a group of learners, a single learner, or no learner at all, the process being the same in any case.

Tutoring may be designed not only to produce mastery of content but also to develop habits of independent inquiry (though it usually is not). Instruction is poorly adapted to this latter function. Nothing is impossible, however, and an eloquent lecturer can--and sometimes does--inspire his listeners to independent

action. But teaching is not teaching if it does not achieve both goals--if it does not both communicate knowledge and develop independent inquiry skills.

In recent years there has been an increasing interest in a third activity, interaction between a teacher and a group of pupils, which is much more commonly seen in a typical classroom than either tutoring or instruction. Ubiquitous though it is, this phenomenon was largely ignored by educational researchers until a few years ago. Nowadays, the study of classroom behavior per se is so popular that there is some danger that research into other aspects of teaching may be neglected. Amidon and Hunter (1966), like many other students of the teaching process, define teaching as interaction. While this is undoubtedly carrying things too far, there is no questioning the importance of research into this aspect of teaching.

The recent increase in research in teaching has been made possible by the development of a growing number of devices for analyzing teaching, known variously as observation schedules, interaction analysis techniques, verbal coding systems, and so on. A detailed review of these developments is outside the scope of the present discussion: the interested reader may consult Gage (1963); Volume 14, No. 3 of the Journal of Teacher Education; or Amidon & Hough (1967), for an introduction to the subject. (An anthology of such instruments may be found in Simon & Boyer, 1967 and 1970.)

Among these devices the present report will be concerned with

only two. One, the technique of Interaction Analysis introduced by Flanders (1960, 1965), is included because it is without question the most widely used and best known method for analyzing teacher-pupil interaction (cf., e.g., Amidon & Hunter, 1966; Amidon & Hough, 1967). It is included in this study as a point of reference to which the second device may be related and against which it may be compared.

This second instrument, the Observation Schedule and Record, Form 4V, or OScAR 4V (Medley, Impelletteri & Smith, 1966), is not altogether new. The first version of OScAR appeared more than a decade ago (Medley & Mitzel, 1958), and it or some variation has been used in numerous studies since (e.g., Fowler, 1962, 1966; Morrison, 1961, 1966, 1967; Bowers & Soar, 1960, 1961; Soar, 1962; Allen, 1969).

The "OScAR technique" was originally conceived more as an approach to the measurement of classroom behavior than as an instrument per se (Medley & Mitzel, 1958, 1963). As an instrument it has been subjected over the years to a constant process of revision of which Form 4V, the one used in this study, was a product. It has been the intent in this process to develop and improve a procedure by which teaching style could be measured objectively and reliably.

The general approach has been inductive. After preliminary work with Withall's categories (Withall, 1948, 1949, 1951, 1956) indicated that modification was in order (Mitzel & Rabinowitz, 1953; Medley & Mitzel, 1955), a simplified set of categories was

developed and used in one study (Medley & Mitzel, 1958, 1959). Over the years from 1959 to 1966 a number of revisions were made and tried out, until the present version was devised and submitted to the relatively extensive empirical tryout reported below.

The primary purpose of the present study was to develop scoring keys for OScAR 4V which would reliably and validly measure dimensions of classroom behavior which are of theoretical and/or practical interest and which are also stable enough to be described as elements of teaching style. Identification of such dimensions and development of procedures for measuring them would seem to be prerequisite to efforts to help teachers develop and improve their effectiveness in the classroom as well as to research into the nature of effective teaching and teacher education.

The names attached to such dimensions are seen as a first attempt to define some of the "words" that would constitute a language of teacher behavior (cf. Medley, 1971). Such dimensions should have three properties: (a) They should be operationally defined in measurable terms; (b) they should be immediately useful in the context of supervision; and (c) they should bear sufficient apparent relationship to current theories of teaching to make their relationship to teacher effectiveness worth studying.

Other techniques, notably the one already mentioned, developed by Flanders and included in this study, possess the latter two properties; but most of them lack the first. It is the authors' conviction that the usefulness of a device in a scientific investigation depends on the degree to which it possesses the essential

characteristics of a measuring instrument, viz., objectivity, reliability and validity.

The word reliability as applied to observational data is usually interpreted as referring to observer agreement. To the extent that records of the same behaviors made by different observers agree, the observations are described as reliable. Evidence of such agreement is better regarded as evidence of objectivity than of reliability, except, perhaps, when the purpose for which the observations are to be used is one which depends on obtaining an accurate description of what happened in that classroom at one particular time--the time when the observations were made. But when the purpose is to measure elements of style which persist over time, and which should be apparent at different times--as is the case with OScAR--this concept of reliability is inadequate. The proper way to define the reliability of observations of teacher style, then, is to say that a dimension is reliable to the extent that scores on the dimension based on records made by different observers observing the same teacher at different times agree. So defined, reliability implies objectivity.

When this definition of reliability has been employed, as has been the case in all earlier work with OScAR (cf. Medley & Mitzel, 1955, 1958, 1963; Mitzel & Medley, 1962), results have consistently indicated that, in comparison with errors of measurement resulting from instability of teacher behavior from day to day, errors due to disagreement of observers on how to code the same behavior are minor--even negligible. Unreliability of observations arising from

lack of objectivity in this type of instrument has been found to be consistently insignificant.

In the present instance, therefore, a calculated risk was taken, and such errors were assumed to be zero. This implies that reliability in this study depends on the extent to which records made by the same observer at **different** times agree. In other words, coefficients of stability will be reported rather than coefficients of reliability, on the grounds that experience has shown that in practice the two differ very little.

The validity of observational data is best thought of as a matter of content validity. Since the behaviors to be measured are inaccessible except through observation, evidence of objectivity, evidence that different observers observing at the same time describe the behavior in the same way--is evidence of validity or (to put it differently) evidence that the descriptions of behavior are accurate.

The assumption that observer disagreement is negligible in this study is tantamount to assuming that the observations are valid; therefore, direct evidence of validity cannot be reported in this study; although the stability coefficients reported may be taken as indirect evidence. In addition, evidence of correlations between OScAR dimensions and other variables: experience, subject taught, scores on the National Teacher Examination, and scores on the Flanders instrument based on observations made by other observers, also constitute indirect evidence of validity.

If two instruments as different in conception and execution

as OScAR and the Flanders' system yield descriptions of the same behavior that are closely similar (in some respects at least) that fact may be taken as evidence of validity in the sense that what they both describe must be there. If relationships can be established between scores on behavior dimensions and such things as amount of experience, subject taught, or how much and what a teacher knows (as indicated by his scores on achievement tests), such a relationship indicates not only that the observations measure actual behaviors, but that they measure aspects of behavior that are important in the sense of being useful. Indeed, it is through such a process that the "words" of the instrument gain meaning.

OScAR 4V

It is difficult to reconstruct the procedure by which the earlier OScARs (which were nothing but extensions of the Withall Categories) became OScAR 4V (a new and different instrument altogether). In retrospect it seems to have been a series of minor breakthroughs or innovations. The first was the realization that the system could be based entirely on the content of teacher utterances--that it was neither necessary nor particularly desirable to record or classify pupil verbal behaviors on any basis except the teacher's reactions to them. This greatly simplified the observer's task, first, by enabling him to concentrate on only one person instead of trying to observe 25 or 30 simultaneously; and, second, by shifting the task of evaluating and interpreting pupil behavior from the observer to the teacher, so that the

observer needed to interpret teacher behaviors only. And this he was able to do by looking at them from the pupil's point of view--by making only such discriminations as a pupil could be expected to make.

A second advance came when the concept of the interchange as a single event, to be coded and recorded as such, was formulated. An interchange is an episode in which (a) a pupil speaks to the teacher, and (b) the teacher is expected to evaluate the pupil's utterance. The most common type of interchange begins with a teacher question; while the pupil thinks up an answer, the observer has time to code the question. By the time the teacher evaluates the answer, the observer is ready to code the evaluation; he then makes one tally for the entire event in a two-way classification scheme. An observer soon develops an effortless rhythm in doing this which makes his task easier and his record more accurate at the same time. The interchange appears to be a natural behavior unit; certainly it seems natural to code it as a single event.

The third important innovation was the distinction made between initiating statements and continuing statements. A statement is defined as an utterance which does not solicit any response. By distinguishing between a continuing statement, which is one classifiable in the same category as the one immediately preceding it, from an initiating statement, which is not, the number of interpretable statement categories was doubled without requiring the coder to memorize any additional categories. More

important, the information obtained about how long a teacher tends to continue in the same vein is very useful. For one thing, it makes it possible to distinguish a lecturer from a teacher who intersperses informing statements between questions but never really lectures. For another, it makes it possible to measure the intensity as well as the amount of affect in a teacher's behavior, since strong affect usually takes more than one statement to be adequately expressed.

Description of the Categories.

Brief definitions of the 21 categories used on OScAR 4V are given in Table 1.

Insert Table 1 about here

Basically, three major categories of behavior are recognized according to content--affective, substantive, and procedural; and two major categories are recognized according to form--statements and interchanges. Statements are classified first on content and then on sequence (as Initiating or Continuing). Interchanges are classified first according to entry--i.e., the manner in which the pupil utterance was elicited; and second according to exit.

Four types of entries to substantive interchanges are recognized, one pupil-initiated, three teacher-initiated. The three teacher-initiated entries or questions--Convergent, Divergent, and Elaborating--are discriminated on the basis of what kind of pupil response they seem to solicit. If, in order to answer a teacher

Table 1

Summary of Categories of Verbal Behavior on OScAR 4V

I. STATEMENTS

A. Teacher statements--utterances which neither respond to nor solicit a response from a pupil--are classified as follows:

1. AFFECTIVE. A statement revealing sensitivity to pupil feelings is classified as CONSIDERING. A statement criticizing pupil conduct is classifed as REBUKING.

2. SUBSTANTIVE. A statement containing no affect but referring directly to content to be learned by pupils is classified as INFORMING if it conveys a fact, generalization, or the like, or PROBLEM STRUCTURING if it sets up a question or issue to be solved.

3. PROCEDURAL. A statement which contains neither affect nor substance is classified as DIRECTIVE if it contains a command or an instruction with the force of a command. A statement which does not clearly fall into one of the above categories is classified as DESCRIBING.

B. Pupil statements--utterances by pupils addressed to other pupils are classified as PUPIL STATEMENTS.

C. Sequence. If a teacher makes two or more successive statements which may be classified in the same category, all except the first are classified as CONTINUING. The first statement in a series of the same kind is classified as INITIATING.

II. INTERCHANGES

An interchange is an episode in which a pupil says something to the teacher and the teacher reacts.

A. Substantive interchanges are those in which the pupil's utterance refers to content to be learned. Such interchanges contain two parts: entry and exit.

1. Entries. A substantive interchange begins with one of four types of entries:

a. PUPIL INITIATED. The pupil addresses a statement or question to the teacher.

b. ELABORATING. The teacher addresses a question to a pupil which refers directly to a previous pupil comment.

c. DIVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment, and which offers him a choice of two or more acceptable or "correct" answers.

d. CONVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment and to which there is only one acceptable answer.

2. Exits from Substantive Interchanges. After the pupil has asked his question or made his answer, the teacher disposes of the answer in one of six ways, called Exits. Exits are first classified according to the information they contain about the correctness or acceptability of what the pupil has said.

If the teacher clearly indicates that what the pupil has said is correct or acceptable, the interchange is classified as SUPPORTED if praise or enthusiasm is shown, as APPROVED if praise is not given.

If the teacher clearly indicates that what the pupil has said is incorrect or unacceptable, the interchange is classified as CRITICIZED if disapproval of either the pupil or what he has said is expressed, or as NEUTRALLY REJECTED if no disapproval is expressed.

If the teacher makes some response to what the pupil says which does not clearly indicate whether it is correct (acceptable) or incorrect (unacceptable) the interchange is classified as ACCEPTED; if the teacher makes no response, it is classified as NOT EVALUATED.

B. Non-Substantive Interchanges are those in which the pupil's contribution does not refer to content to be learned.

1. TEACHER-INITIATED non-substantive interchanges are classified as POSITIVE or NEGATIVE according to the affective content of the teacher's question.

2. PUPIL-INITIATED non-substantive interchanges are classified as POSITIVE if the teacher supports, approves, or accepts the pupil's suggestion, and as NEGATIVE if he criticizes, neutrally rejects, or ignores it.

question, the pupil addressed must have attended to the pupil utterance which preceded the question, the entry is classified as Elaborating. If not, and if it appears that a particular answer is solicited, and is the only one that will be accepted, the entry is classified as Convergent; otherwise the entry is coded as Divergent.

The definitions of these last four categories exemplify the kind of discrimination the OScAR technique requires. The cues used in coding these entries are simple, overtly observable, and are not based on value judgments. They are easy to learn, and require no psychological or pedagogical sophistication of the observer. And yet they seem to get at important aspects of interaction--whether the teacher asks more closed or open questions, whether questions are interrelated, whether pupils are expected to listen to one another, etc.

Exits from substantive interchanges are classified into six categories. It is assumed that whenever a pupil says something to a teacher about the substantive content of the lesson, the expectation is that the teacher will either label that item of content as correct (or acceptable) or label it as incorrect (or unacceptable). Which of the two he does, or whether he does neither, is the first rubric on which exits are classified--positive feedback, negative feedback, or no feedback. In addition, the teacher may add praise (when feedback is positive) or criticism (when it is negative). When no feedback is given, the pupil's utterance may or may not be overtly acknowledged.

These six categories may be regarded as resembling six regions lying along a continuum in the following order: Supported, Approved, Acknowledged, Not Evaluated, Neutrally Rejected, Criticized. Accordingly, behaviors codeable in adjacent categories may be difficult to discriminate in borderline cases. The instruction given is that if an exit is near the dividing line between two categories it should be coded in the "inside" category--e.g., if the observer is not quite certain a pupil was Supported, he should classify the interchange as Approved, because the latter category lies nearer the center of the continuum.

Non-substantive interchanges (as they are called) are not really interchanges at all, since no pupil utterance is offered for teacher evaluation. They are classified first according to origin--Teacher-Initiated or Pupil-Initiated; and then according to the content of the teacher utterance. A Pupil-Initiated Non-Substantive Interchange usually is an attempt to initiate procedure (a request for permission, for instance) and is classified as Positive if the teacher responds in a manner indicating that the initiation was acceptable--if the teacher gives the permission asked, declines with regret, etc. If the teacher does not indicate that the initiative was acceptable--refuses the permission, rebukes the pupil, etc., it is coded Negative.

When the teacher offers a pupil a chance to initiate procedure, or at least a choice of action; or when she asks a non-substantive, supportive question (Do you feel all right?), the interchange is Teacher-Initiated, Non-Substantive, Positive.

Negative Teacher-Initiated Non-Substantive Interchanges are uncommon; they occur when a teacher rebukes a pupil with a question to which he is expected to make a verbal response (Did I give you permission to leave your seat?).

A non-interactive teacher statement is examined first for affective content. If it contains positive affect or awareness of pupil feelings, it is classified as Considering; if it contains negative affect, it is classified as Rebuking. Statements that are emotionally neutral are examined for substantive content. If a statement contains information which pupils are supposed to learn and remember, it is classified as Informing. If it raises a question about content, but does not solicit a response from a pupil, it is called Problem Structuring. And statements containing neither affect nor substance are coded Directing if they restrict or channel pupil behavior, Describing if they do not.

Pupil-Initiated utterances which do not solicit teacher response are classified as Pupil Statements.

As has been mentioned, a statement which falls in the same category as the one immediately preceding it is classified as a Continuing Statement in that category; one which falls in a different category than the one preceding it is classified as an Initiating Statement in the category in question.

To sum up: the observer using OScAR 4V must learn 19 categories for teacher utterances and three for pupil utterances; but his observations are recorded in such a way as to yield frequencies in 42 different categories. Each tally on a record

represents one event. An event is either a non-interactive one in which only one person speaks; or an interchange, in which a pupil speaks once and the teacher may speak once or twice.

The form used by the observer in recording teacher behavior is shown in Figure 1.

Insert Figure 1 about here

The OScAR coder makes a tally in the appropriate space on the form for each event he observes; a complete record shows the number of events of each type recorded during one visit to the classroom.

Scoring Procedures for OScAR 4V

A raw OScAR 4V record consists, then, of a vector or profile of 42 frequency counts corresponding to the 42 events discriminated by the procedure. The relative frequency of each type of event in the profile of a teacher is the primary datum on which a description of that teacher's teaching style must be based.

The objectivity of the record depends, of course, on the degree to which such frequencies reflect teacher behavior rather than observer idiosyncrasy. As has been mentioned, considerable effort was expended in devising categories in such a way as to ensure objectivity in the records. While categories were also designed to relate to elements of teaching style that appeared to be psychologically or pedagogically important and meaningful, objectivity was the first consideration. As a result, the psychological or pedagogical meaning of a single frequency count tends

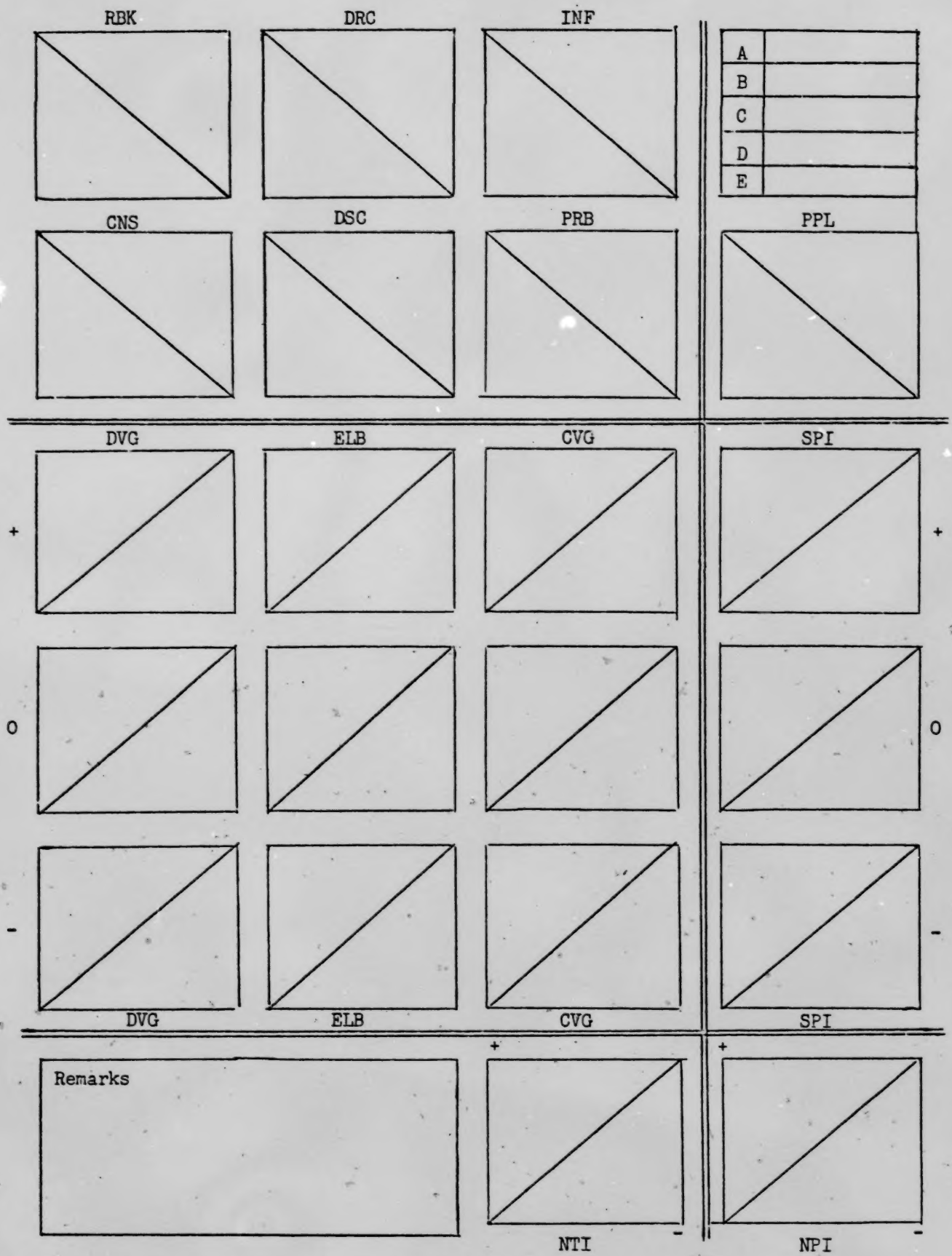


Figure 1. OScAR LV

to be somewhat ambiguous. Different teachers can run up similar frequencies in a given category in quite different ways. How often a teacher criticizes a pupil's response to a question, for instance, may depend on how high the teacher's standards of acceptable behavior are, on how often his pupils give incorrect answers, on how much latent hostility the teacher possesses, on how important he thinks it is to give negative feedback, or on something quite different from any of these.

There are two ways of distinguishing between teachers who get equally high scores on a single category for different reasons. One possibility is to subdivide the category in such a way that they will not get identical scores--to recognize several different rebukes, for example. This approach usually results in a larger, more difficult, and less reliable set of categories. The other approach is to look at other items in conjunction with the one in question. There may be another category on which teachers can also score high for various reasons, only one of which is shared with the original category. The <u>total</u> frequencies of the two teachers across both categories will then differ--one will score high on both, the other will not.

Suppose, for instance, that the two categories in question are <u>Convergent Criticized Interchanges</u> and <u>Rebuking Continuing Statements</u>. A teacher who criticizes pupil answers because of his own latent hostility will score high on both items; but one who criticizes pupil answers for some other reason--because his standards are high, his questions are difficult, etc.--will score

high only on the first category. Similarly, a teacher who rebukes his pupils often because they misbehave frequently will score high on the second category but not on the first. Only a teacher who rebukes and criticizes pupils because of his own hostility will score high on both. Thus a composite score on both items is a more valid measure of teacher hostility than a single score on either item. As two, three, or more items are combined for scoring into "keys" (as they are called), the scores become less ambiguous (and consequently more valid as descriptions of teacher behavior) without loss of objectivity.

In this way it becomes possible for the instrument to make discriminations and to measure variables more subtle and sophisticated than those the observers need to recognize, and at the same time to preserve in the measurements a level of objectivity possible only when observer discriminations are relatively simple and obvious. It is as though children were sent as observers to a zoo to find out what kinds of animals were there, and instead of being asked to identify each animal, they were asked to discriminate hooves from claws, fur from feathers, horns from antlers, etc., and to record which characterized each animal they observed. From such records it would be possible to infer accurately and reliably what species of animals inhabited the zoo, and to recognize descriptions of animals the children had never heard of and could not, perhaps, themselves identify. This is the central idea underlying the OScAR technique--the idea of basing complex dimensions on elementary discriminations (cf. Medley, 1970).

When there are only 42 categories to work with (as is the case with OScAR 4V), the number of multi-item variables or keys which can be developed before the problem of item overlap becomes acute is rather small. If two keys, each of which contains a relatively small number of items, have one or two items in common, there will be a substantial lack of experimental independence between them. This means that chance variation on an item shared by both keys will contribute correlated errors of measurement to scores on both keys, and produce spurious correlations between them. The use of conventional procedures for interpreting test scores (significance tests, confidence intervals) assumes that each bit of chance variation will appear only once, and is designed to guard against misinterpretation resulting from such variation. Use of these methods with overlapping keys could cause severe overinterpretation of chance differences and lead to erroneous conclusions.

It seemed important to ensure as far as possible that each score obtained from OScAR 4V should be experimentally independent of each other score. The degree to which this could be achieved depends,on the degree to which the 42 category frequencies from which the scores must be derived are themselves independent. The problem is to derive mutually independent combinations of category frequencies.

The easiest way to do this is, of course, to use non-overlapping keys. If one key is a total over categories A and B, and another over C and D, the keys do not overlap and there is no

problem. However, if a third score equal to the total frequency over all four categories is also employed, the condition is not fulfilled, since this third key is dependent on both of the others.

What seems to be needed here is information about the total frequency over all four categories--A, B, C, and D--and information about the difference in frequency on A and B and that on C and D.

A Note on Orthogonal Contrasts

The use of orthogonal **contrasts** as employed in the analysis of variance (Li, 1964; Winer, 1962) seems to provide a solution to this problem. Following this method, two keys would be used. One would be the total over all four categories already suggested. The other would be the difference between the total of A and B and the total of C and D. The scoring weights could be represented as shown in Table 2.

Insert Table 2 about here

Suppose that we generate sets of four numbers randomly and assign them randomly as values of frequencies A, B, C, and D. It should be apparent that the expected correlation between A and B (or any other pair) would be zero. In other words, A, B, C, and D are experimentally independent. Also, A + B is independent of C + D in the same sense; but A + B is not independent of A+B+C+D because of the overlap between them. But, (A + B) - (C + D) is

experimentally independent of A+B+C+D, despite their overlap. For any given value of the one, any value of the other is possible.

These two keys are orthogonal--that is, to the degree that A, B, C, and D are experimentally independent, Key 1 and Key 2 are also independent. The condition proposed--that no dependency be introduced by adopting the new keys--is fulfilled despite the complete overlap of items on the two keys. Scores on the two keys may be interpreted in the usual ways without conclusions based on one being distorted by results on the other.

Using this approach, and choosing for the first key the simple total of K available categories, it is always possible to develop K-1 additional keys all of which are orthogonal to the first and to each other. In fact, this may be done in more than one way--more than one set of K-1 contrasts (as they will be called) may be constructed. The only restriction on the weights is that they must sum to zero. Only one set may properly be used, however, since contrasts from different sets will not be orthogonal. The trick is to select a set such that each key makes the best sense, psychologically or pedagogically.

The criterion for recognizing whether two keys are orthogonal to one another is that the sum of cross-products of scoring weights for the two keys over all the items must be zero. In the example above, for instance, this sum is $(-1)(+1) + (-1)(+1) + (+1)(+1) + (+1)(+1) = (-1) + (-1) + (+1) + (+1) = 0$, which satisfies the criterion.

Note that the original proposal to use the total of A and B,

the total of C and D, and the total of all four as keys, would give the weights represented in Table 3.

Insert Table 3 about here

Keys X and Y would be independent of each other, since the sum of cross-products of their weights is zero; but neither would be orthogonal to Key 1, since the sum of cross-products in either case would be +2.

In order to illustrate the procedure used in selecting a set of orthogonal keys, it will be assumed that categories A, B, C, and D represent the OScAR categories indicated in Table 4.

Insert Table 4 about here

Key 1 may now be recognized as representing the total number of affective teacher statements tallied on a record. It may be said to measure the amount of affect observed. Key X (in Table 3) is recognized as the total number of statements showing negative affect and Key Y the total number showing positive affect.

Key 2 (in Table 4) is seem to be the difference between the numbers of positive and negative affective statements on a record--a measure of the direction of affect experimentally independent of its amount (as measured by Key 1).

One way in which two additional independent keys to use with Keys 1 and 2 might be selected is shown in Table 4.

Table 2

Weights Used On Two Hypothetical Orthogonal Scoring Keys

Category	A	B	C	D
Key 1	+1	+1	+1	+1
Key 2	-1	-1	+1	+1

Table 3

Hypothetical Keys Not Orthogonal to Each Other

Category	A	B	C	D
Key 1	+1	+1	+1	+1
Key X	+1	+1	0	0
Key Y	0	0	+1	+1

Table 4

A Set of Four Orthogonal

Keys Scored on Four OScAR Items

Category	Rebuking Statements Initiating	Rebuking Statements Continuing	Considering Statements Initiating	Considering Statements Continuing
	(A)	(B)	(C)	(D)
Key 1	+1	+1	+1	+1
Key 2	-1	-1	+1	+1
Key V	-1	+1	-1	+1
Key W	-1	+1	+1	-1

Table 5

Weights Used to Score Contrasts Among

Four Affective Categories on OScAR 4V

Category	Rebuking Statements Initiating	Rebuking Statements Continuing	Considering Statements Initiating	Considering Statements Continuing
Key 1	+1	+1	+1	+1
Key 2	-1	-1	+1	+1
Key 3	0	0	-1	+1
Key 4	-1	+1	0	0

Key V contrasts continuing affective statements with initiating ones, ignoring direction. It would seem to reflect intensity or depth of affect. When a teacher makes several statements in succession all of which are classified as Rebuking, he is usually rebuking someone severely; when the rebuke consists of only one sentence it is generally less intense or severe. A teacher whose average number of statements per rebuke is relatively high, then, is probably showing more hostility, more rancor, or at least making a bigger thing of his rebukes than one whose average number of statements per rebuke is near the minimum possible value of one.

A similar line of reasoning applies to Considering Statements--a longer series is less likely to be perfunctory and more likely to be strongly supportive than a single statement in this category.

Key V, then, should discriminate teachers whose affective statements (positive or negative) are perfunctory from those who tend to express deeper or strong affect. And it should do so independently of the amount and direction of affect (Keys 1 and 2).

Key W, the fourth of the set, gives positive weights to perfunctory consideration and intensive rebukes, and negative weights to perfunctory rebukes and intensive consideration. This key should contrast teachers who act as if they actively dislike pupils with those who seem genuinely to like them. Again, this is independent of the amount, direction, or depth of affect.

The reader may verify the orthogonality of these keys by working out all six sums of products of weights for himself. This

set of keys will illustrate what can be done with orthogonal contrasts; how their use makes it possible to derive independent combinations of items which reflect different a priori hypotheses about how behavior is organized. Scoring a contrast like Key V (for example) amounts to proposing an hypothesis that there is a dimension of teaching style which manifests itself in the indicated behavior pattern. Asserting that a behavior dimension exists implies that reliable differences between teachers will be found on the dimension. The hypothesis may be tested by estimating the reliability of scores on that dimension; if they are reliable, the dimension exists.

The set of three contrasts just proposed will be recognized as the set obtained when the four categories on which they are based are regarded as forming a two-by-two factorial design. Key 2 is the main effect for direction of affect, Key V the main effect for intensity, and Key W the one for interaction.

In the study being reported, a different set of keys was used as shown in Table 5.

Insert Table 5 about here

This set will be recognized as based on a hierarchical design (between and within). Keys 1 and 2 are as before; Key 3 is sensitive to intensity of positive affect, Key 4 to intensity of negative affect. By measuring intensity separately in the two types of affect, access to the interaction is lost; but at the time at

which the decision was made it was felt that this set of keys was more likely to yield meaningful dimensions than the set in Table 4.

The Set of Contrast Keys

The number of different sets of 41 contrasts that could be set up among the 42 categories on OScAR 4V is very large. The approach used in selecting the set finally adopted was to begin by dividing the categories into subgroups. A total score on each subgroup was used together with a set of contrasts within that subgroup of categories.

One subgroup consisted of the four affect-related statement categories already described, and the contrasts used are those already shown in Table 5, which also appear as Keys 1 to 4 in Table 6, under the respective titles: (a) Affective Statements, (b) Positive Affect, (c) Concern, and (d) Hostility.

Insert Table 6 about here

A second set consisted of two keys based on Pupil Statements, shown in Table 6 as Keys 5 and 6. One (Pupil Talk) reflects the total number of Pupil Statements recorded, and the other (Continuing Pupil Talk) contrasts the numbers of Continuing and Initiating Statements, to indicate how long (on the average) a pupil is permitted to hold the floor when he speaks on his own initiative.

A third set included the four statement categories related to procedure and the four categories of non-substantive interchanges.

Table 6

Keys Based on Statements and Non-Substantive Interchanges

Key No.	Name	Statement Categories: Rebuking Init	Rebuking Cont	Considering Init	Considering Cont	Directing Init	Directing Cont	Describing Init	Describing Cont	Informing Init	Informing Cont	Problem Structuring Init	Problem Structuring Cont	Pupil Statements Init	Pupil Statements Cont	Teacher Initiated Pos	Teacher Initiated Neg	Pupil Initiated Pos	Pupil Initiated Neg
1.	Affective Statements	+1	+1	+1	+1														
2.	Positive Affect	-1	-1	+1	+1														
3.	Concern			-1	+1														
4.	Hostility	-1	+1																
5.	Pupil Talk													+1	+1				
6.	Continuing Pupil Talk													-1	+1				
7.	Management					+1	+1	+1	+1							+1	+1	+1	+1
8.	Control					+1	+1	+1	+1							-1	-1	-1	-1
9.	Dominance					+1	+1	-1	-1										
10.	Direction Complexity					-1	+1												
11.	Description Complexity							-1	+1										
12.	Permissiveness															+1	-1	+1	-1
13.	Pupil Initiative															-1	-1	+1	+1
14.	Benevolent Despotism															+1	-1	-1	+1
16.	Suspense									-1	-1	+1	+1						
17.	Problem Complexity											-1	+1						
18.	Lecturing									-1	+1								

The weights assigned to these eight categories are shown in Table 6, Keys 7-14.

Key 7 (Management) indicates the total number of events--statements and interchanges--devoted to procedural matters. Key 8 (Control) identifies the teacher who keeps the initiative in procedural matters tightly held in his own hands, neither offering pupils any chance to discuss what they will do nor allowing them to raise questions about what he decides they are to do. Key 9 (Dominance) indicates whether a teacher discusses or explains what the class is doing or simply gives orders. Key 10 (Direction Complexity) indicates how detailed or elaborate a teacher's commands are, and Key 11 (Description Complexity) how extensive his explanations or comments about procedure are. Number 12 (Permissiveness) contrasts the teacher who frequently offers pupils a choice of actions, and usually gives permission when it is asked for, with the teacher who rarely or never does either. Key 13 (Pupil Initiative) indicates the relative number of times pupils attempt to initiate procedure compared to the number of times they are asked to do so. And a high score on key number 14 in Table 6 would be earned by a teacher who often offers pupils a choice, but seldom gives them permission to do something when they initiate the request--a Benevolent Despot.

The fourth group of keys has to do with substantive matters--with verbal behaviors related to subject-matter content. It includes four statement categories and 24 categories for substantive interchanges. The first key in this group is not shown in

Table 6; it contrasts the number of substantive teacher statements with the number of substantive interchanges. It seems to indicate the degree to which content is developed by teacher statements as against teacher-pupil interaction. It is key number 15, called Teacher Talk.

The other three keys are numbers 16-18 in Table 6. Key 16 (Suspense) discriminates teachers who tend to raise questions about content before giving information about content from teachers who merely inform. Key 17 (Problem Complexity) indicates how many statements a teacher uses on the average in structuring a problem. Key 18 (Lecturing) indicates how many Informing Statements in a row the teacher makes. The "lecturer" tends to talk at length--hence he accumulates Continuing Informing Statements in his record.

There should have been a key based on a simple total of all 28 content-related behaviors--Substantive Statements and Substantive Interchanges; but through some oversight, a key based only on all 28 categories of "interchanges" was scored instead. Scores obtained on this key were not orthogonal to scores obtained on Keys 7 and 15, so it was omitted from the study.

The set of 23 contrasts scored on the 24 Substantive Interchange categories included three contrasts based on the four types of entries, five contrasts based on the six types of exists, and 15 based on interactions between the two.

Table 7 shows the three keys based on the four entries to substantive interchanges, which were obtained by partitioning the

Insert Table 7 about here

three degrees of freedom among them into a three-level hierarchy. Key 19 (Teacher Initiative) contrasts interchanges beginning with teacher questions with ones beginning with pupil questions (or statements). Key 20 (Cohesion) contrasts interchanges beginning with elaborating questions--i.e., questions requiring a pupil to react to a preceding pupil response--with interchanges beginning with questions which do not directly relate to preceding answers. Key 21 (Divergence) contrasts interchanges beginning with questions to which the pupil has a choice of acceptable answers with interchanges beginning with questions calling for predefined answers.

Table 8 shows the five keys based on exits. Key 22 (Feedback)

Insert Table 8 about here

contrasts those interchanges in which the teacher explicitly informs a pupil whether his comment was acceptable (correct) or unacceptable (incorrect) with those interchanges in which the pupil is not so informed.

Key 23 (Valence) contrasts positive and negative feedback, and indicates whether (as judged by the teacher) pupils are right more often than they are wrong.

Key 24 (Enthusiasm) is designed to discriminate those teachers whose feedback lacks affect from those who tend both to praise

Table 7

Keys Based on Entries to Substantive Interchanges

Number of Contrast	Name of Key	Type of Question			
		Pupil Initiated	Divergent	Elaborating	Converge
19.	Teacher Initiative	-1	+1/3	+1/3	+1/3
20.	Cohesion	0	-1/2	+1	-1/2
21	Divergence	0	+1	0	-1

Table 8

Keys Based on Exits From Substantive Interchanges

Number of Contrast	Name of Key	Exit					
		Supporting	Approving	Acknowledging	Not Evaluating	Neutrally Rejecting	Criticizing
22.	Feedback	+1/2	+1/2	-1	-1	+1/2	+1/2
23.	Valence	+1	+1	0	0	-1	-1
24.	Enthusiasm	+1	-1	0	0	-1	+1
25.	Encouragement	+1/2	-1/2	-1	+1	+1/2	-1/2
26.	Positivity	+1	-1	+1	-1	+1	-1

acceptable and to criticize unacceptable pupil responses.

Key 25 (Encouragement) and Key 26 (Positivity) are both designed to get at the same characteristic--a tendency for the teacher to "shift" his evaluation toward the supportive side--but are based on different hypotheses about what happens when the teacher does not give feedback. One hypothesis is that when a pupil response is incorrect, this type of teacher does not like to acknowledge it (because if he acknowledged it the pupil might think his response was acceptable) and therefore keeps quiet instead. Therefore, key 25, Encouragement, weights acknowledging exits negatively and weights interchanges not evaluated at all positively. An alternative hypothesis is that this type of teacher feels that failure to acknowledge a pupil's response (when it is unacceptable) may discourage him. So key 26, labeled Positivity, weights acknowledging exits positively and those not evaluated positively. (Both keys give positive weights to interchanges in which the teacher praises correct responses and ones in which he refrains from criticizing incorrect ones.)

The remaining set of keys, based on the 15 "interaction" contrasts, are listed in Table 9, but the scoring weights are not

---

Insert Table 9 about here

---

shown. Each possible combination generated by combining any one of the five exit contrasts with one of the three entry contrasts provides one key. Table 10 shows as an example, the weights for

Table 9

OScAR Entry and Exit Keys Based on Interaction

Number	Exit	Entry
27	Feedback	Teacher Initiative
28	Feedback	Cohesion
29	Feedback	Divergence
30	Valence	Teacher Initiative
31	Valence	Cohesion
32	Valence	Divergence
33	Enthusiasm	Teacher Initiative
34	Enthusiasm	Cohesion
35	Enthusiasm	Divergence
36	Encouragement	Teacher Initiative
37	Encouragement	Cohesive
38	Encouragement	Divergence
39	Positivity	Teacher Initiative
40	Positivity	Cohesion
41	Positivity	Divergence

Insert Table 10 about here

the key obtained by combining 19, Teacher Initiative, and 23, Feedback. Note that the cell weights are products of marginal weights from "main effects" keys, so that the reader may, if he like, calculate the weights on any key for himself. A teacher who scored high on Key 27 would be one who tends to give more positive feedback to pupil comments which are responses to questions than he gives to unsolicited or voluntary comments, and one who tends to give negative feedback oftener on voluntary contributions than on ones he has solicited directly.

It should be remembered that this key is independent both of Key 23 (which measures the total amount of positive and negative feedback the teacher gives) and of Key 19 (which measures the relative number of pupil-initiated interchanges). This makes it possible by intercorrelating the keys to find out whether (as might be expected) the teacher who rewards pupils who do not volunteer information is the one whose pupils volunteer least often. It also makes it possible to identify classrooms (if they exist) in which, despite the fact that the teacher continually rejects unsolicited pupil contributions, the pupils continue to make them; and it provides a means of trying to find out why. It is this kind of capability that makes orthogonality of scoring keys so desirable.

Table 10

Weights Assigned to 24 Types of Interchanges on

Feedback by Teacher Initiative

Exit	Entry				Key 22 Feedback
	Pupil Initiated	Divergent	Elaborating	Convergent	
Supported	-1/2	+1/6	+1/6	+1/6	(+1/2)
Approved	-1/2	+1/6	+1/6	+1/6	(+1/2)
Acknowledged	+1	-1/3	-1/3	-1/3	(-1)
Not Evaluated	+1	-1/3	-1/3	-1/3	(-1)
Neutrally Rejected	-1/2	+1/6	+1/6	+1/6	(+1/2)
Criticized	-1/2	+1/6	+1/6	+1/6	(+1/2)
(Key 19. Teacher Initiative)	(-1)	(+1/3)	(+1/3)	(+1/3)	

Flanders' Interaction Analysis Technique (FIAT)

The technique of interaction analysis introduced by Flanders in 1970 (Flanders, 1960, 1965) is probably the best known and most widely used observation instrument in existance, and there is a considerable literature in print about it (cf. Amidon & Hough, 1967; Flanders, 1970), so it will be described here only briefly. The observer using the technique memorizes the numbers and definitions of the 10 categories shown in Table 11. During each visit to a classroom his basic task is to write down every three seconds

---

Insert Table 11 about here

---

a number between one and ten which best characterizes the verbal behavior observed during the preceding three-second interval. When the behavior shifts from one category to another during a single three-second interval, two numbers are recorded (in the appropriate order) instead of one. A complete record for one visit is, then, a string of numbers between one and ten, preserving a sequential description of the kinds of verbal behavior heard by the observer during that visit.

Before any attempt is made to interpret a record made by Flanders' Interaction Analysis Technique (hereafter referred to as FIAT), the record is transcribed onto a form called a matrix (see Figure 2), consisting of a 10 x 10 array of cells, numbered

---

Insert Figure 2 about here

Table 11

Summary of Categories Used In

Flanders' Interaction Analysis Techniques

1. Accepts Feeling: accepts and clarifies the feeling tone of the students in a non-threatening manner. Feelings may be positive or negative. Predicting or recalling feelings are included.

2. Praises or Encourages: praises or encourages student action or behavior. Jokes that release tension, not at the expense of another individual, nodding head or saying "um hm?" or "go on" are included.

3. Accepts or Uses Ideas of Student: clarifying, building, or developing ideas or suggestions by a student. As teacher brings more of his ideas into play, shift to category five.

4. Asks Questions: asking a question about content or procedure with the intent that a student answer.

5. Lecturing: giving facts or opinions about content or procedure; expressing his own ideas, asking rhetorical questions.

6. Giving Directions: directions, commands, or orders to which a student is expected to comply.

7. Criticizing or Justifying Authority: statements intended to change student behavior from non-acceptable to acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing; extreme self-reference.

8. Student Talk--Response: talk by students in response to teacher. Teacher initiates the contact or solicits student statement.

9. Student Talk--Initiation: talk by students which they initiate. If "calling on" student is only to indicate who may talk next, observer must decide whether student wanted to talk. If he di, use this category.

10. Silence or Confusion: pauses, short periods of silence and periods of confusion in which communication cannot be understood by the observer.

	1	2	3	4	5	6	7	8	9	10	
1											
2											
3					1						
4				1				1			
5					1					1	
6											
7											
8			1								
9											
10					1						Matrix Total
TOTAL			1	1	3			1		1	7
%			14.3	14.3	42.8			143		143	

Figure 2. Sample Interaction Analysis Matrix

from one to ten on both axes. As an example, consider the following brief imaginary record:

5, 5, 4, 8, 3, 5

First a 10 is arbitrarily added at the beginning and end (unless the record happens to begin or end with a 10) so it reads like this:

10, 5, 5, 4, 8, 3, 5, 10

Then each pair of numbers in turn is tallied into the corresponding cell in the matrix, beginning with 10-5, continuing with 55, then 54, etc., so that each number (except the first and last 10) becomes part of two pairs--appearing as the second number in one pair and the first number in a following pair. As shown in Figure 2, this series would be represented by seven tallies in the matrix--one more than the number of numbers in the original record. The set of 100 frequencies tallied in the 100 cells of the matrix may be regarded as the raw data derived from the behavior recorded in the classroom.

The nature of the tallying procedure introduces some difficulties into statistical analyses of cell frequencies, since successive tallies are not experimentally independent. Even if it is assumed that the probability that any given number in a series will equal a given value c ($0 \leq c \leq 10$) is the same regardless of the value of the number immediately preceding it, this is so. For if one pair is (say) x, y, then the probability that the first digit of the pair following it will be y is equal to one. Otherwise the probability would be equal to .10. Also, the probability

that this digit will equal any of the other nine possible values is zero, instead of .10. This dependency would appear to introduce a spurious correlation between cell frequencies. Frequencies in cells 48 and 84, for example, should have a spuriously high intercorrelation, while frequencies in cell 48 and 94 should have a spuriously low one.

This inter-cell dependence was eliminated in this project by a departure from a recommended tallying procedure. Each number in each record obtained in this study was tallied into the matrix once only. Accordingly, the series above would yield four tallies one each in cells 10-5, 54, 83, and 5-10 instead of the seven tallies shown in Figure 2.

The manner in which FIAT data are used imply that the parameters of interest are expected values of percentages of total tallies recorded in each cell (or in certain groups of cells) taken over some population of occasions and observers. It seems obvious that the expected values for individual cells or group of cells will be the same in such a "half-matrix" as in a matrix constructed in the usual way. Estimates from the "full" matrix may be more precise since they are based on larger numbers of tallies, although it is not easy to see just how much more precise they would be. Expected values of total numbers of tallies in each of the ten basic categories, on the other hand, would be estimated just as accurately either way.

Preliminary Analysis of Data

Collection of Data

10. The entire sample of teachers studied was drawn from students in an intern program for secondary school teachers. All students had earned a bachelor's degree before they entered the program. After intensive training during the summer, each one accepted full-time responsibility for a class in a secondary school in Philadelphia or one of its suburbs. They were visited once a week by a college supervisor and also attended an evening seminar weekly during their first year of teaching.

Twelve members of the university faculty regularly assigned as supervisors of interns were used as observers in the study. Six of them were trained to use OScAR 4V, and six to use FIAT. The six observers assigned to OScAR were assembled for three two-hour training sessions spaced over a one-week period. Once the categories had been defined, most of the remaining time was spent in coding practice. Kinescope films of teachers made in elementary school classrooms provided behavior to be coded. At first, the film was stopped every minute or two to allow the coders to compare notes. Between sessions the trainees took OScAR forms along on their regular supervisory visits and practiced with them on the job. By the third session it was seldom necessary to interrupt the film, and it appeared that all observers had control of the system.

The six observers assigned to the Flanders' system received approximately 12 hours of training in the system before the first observations were made.

A sample of 72 interns was chosen and divided into 12 sets of

six teachers, in such a way that the teachers in any one set all taught the same subject at the same level (i.e., either junior or senior high school). Six groups of 12 teachers each were then made up with two sets of six teachers in each group representing two different subjects but only one level. One OScAR observer and one Flanders' observer were assigned to each group of 12 teachers and made all of the observations of these teachers.

The two observers assigned to each group operated as a team, visiting each of the 12 teachers together and recording his behavior simultaneously. On each visit the observers recorded behavior until the team member using FIAT had recorded 560 tallies. Both observers then stopped recording. The average time per visit was, therefore, about 28 minutes.

Two visits were made to each teacher at the beginning of the second semester of the school year, approximately one week apart. Near the end of the semester, after a brief refresher training session, two more visits a week apart were made to each teacher by the same team of observers. Two of the 72 teachers resigned before the study was completed and so were lost to it. The distribution among subjects, levels, and observer teams of the 70 teachers on whom complete data were obtained is shown in Table 12. The raw

---

Insert Table 12 about here

---

data of the study consisted, then, of 280 OScAR records (based on four visits to each of 70 teachers) and 280 records obtained by

Table 12

Composition of The Teacher Sample

Observer Team	Level	Subject Number 1	Subject Number 2	Number of Teachers
1	Junior High	English (6)	Math (6)	12
2	Junior High	English (6)	Science (5)	11
3	Junior High	Social Studies (6)	Science (5)	11
4	Senior High	English (6)	Math (6)	12
5	Senior High	English (7)	Science (5)	12
6	Senior High	Social Studies (6)	Science (6)	12

FIAT.

Among the 70 teachers were 53 who had taken the Common Examinations of the National Teacher Examinations before the commencement of the study. Their answer sheets were secured from the program administration staff (along with those of 36 other interns not observed in the study).

Preliminary Analyses of OScAR Data

A raw OScAR record consists of 42 frequency counts, one for each of the 42 cells; after scoring, the record consists of 42 contrast scores which are linear functions of these frequencies. As a first step in the analysis of the results of the study, a reliability coefficient was calculated for each frequency and for each contrast. Each reliability was estimated from an analysis of variance design following a general procedure described elsewhere (Medley & Mitzel, 1963, pp. 309-321).

Because it plays a central part in a number of different phases of the analysis of data of the study, the specific design used will be described in some detail here. Reference to Table 12, which shows the composition of the teacher sample, may be helpful in understanding what follows. It is perhaps simplest to regard the study as made of six separate but similar studies, one per observer. Observer 1, for example, saw 12 junior high school teachers, six of English, and six of mathematics.

If the records made by observer 1 were analyzed by themselves, the analysis would take the form shown in Table 13. This may be described as basically a two-way design, 4 x 12, without replication.

Insert Table 13 about here

The first dimension is partitioned into three components: Change, contrasting the first two visits (made in January) with the last two (made in May); Visit, contrasting the two first visits of each pair (one made in January and one in May) with the two second visits; and Interaction (Change by Visit). Each of these components is estimated with one degree of freedom, as indicated at Lines 1-3 of Table 13.

The second dimension is partitioned into two components: Subject, shown in Line 4, with one degree of freedom; and Teacher (Within Subject), shown in Line 8, with 10 degrees of freedom. The interaction between the two dimensions is partitioned into six components--those involving Subject being shown in Lines 5-7 and those involving Teacher (Within Subject) being shown in Lines 9-11.

Teacher (Within Subject) and Visit were regarded as random; Change and Subject as fixed. In other words, error terms were selected in such a way that allowance was made for what might have been observed on other occasions in the same classrooms or in classrooms of other teachers of the same two subjects, at the same times of the year as these four observations were made.

In the major analysis, data were pooled across all six groups of teachers in the composite design shown in Table 14. Change,

Insert Table 14 about here

Table 13

Partition of Degrees of Freedom Among Records Made by Observer One

Source of Variation	Degrees of Freedom
1. Change (between visits one semester apart)	1
2. Visit (between visits within one week)	1
3. Interaction, Change by Visit	1
4. Subject (between teachers of different subjects)	1
5. Interaction, Change by Subject	1
6. Interaction, Visit by Subject	1
7. Interaction, Change by Visit by Subject	1
8. Teacher (within subject)	10
9. Interaction, Change by Teacher	10
10. Interaction, Visit by Teacher	10
11. Residual	10
12. Total Variation	47

Table 14

Design for Analyzing Observation

Records of 70 Teachers on Four Visits

	Source of Variation	Degrees of Freedom
1.	Change	1
2.	Visit	1
3.	Change by Visit	1
4.	Subject (Within Observer)	6
5.	Interaction, Change by Subject	6
6.	Interaction, Visit by Subject	6
7.	Interaction, Change by Visit by Subject	6
8.	Teacher (Within Observer and Subject)	58
9.	Interaction, Change by Teacher	58
10.	Interaction, Visit by Teacher	58
11.	Residual	58
	Subtotal	259
12.	Between Observer Variation	20
	Total Variation	279

Visit, and Interaction (Change by Visit) were all estimated across observers; Subject, Teacher, and interactions of such differences with Change, Visit, and the Change by Visit Interaction were estimated within observers and pooled across observers. Since observer differences were confounded with teacher and subject differences, no attempt was made to interpret them; instead they were pooled and are shown as Between Observer Variation.

For purposes of reliability (i.e., stability) estimation, Lines 8, 10, and 11 of Table 4 are of particular interest. The mean square of Teacher (Within Observer and Subject) estimates variance between teachers of the same subject. The mean square for Interaction, Visit by Teacher estimates variation in the behavior of the same teacher between two visits a week or so apart. The mean square for Residual estimates any unexplained variation in teacher behavior.

Table 15 shows a sample analysis of variance, in which scores

---

Insert Table 15 about here

---

on Contrast 2, Positive Affect, were used. The first step in estimating the stability of teacher scores on a key is to compare the mean squares for Interaction, Visit by Teacher (Line 10) and Residual (Line 11) to see whether the former exceeds the latter at a statistically significant level. In this instance, the former was smaller. When this was the case, or whenever the F ratio in question was neither statistically significant nor greater than 2, the two mean squares were pooled for the sake of parsimony (cf.

Table 15

Reliability Analysis of Variance of Scores

on Key 2, Positive Affect

	Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
1.	Change	1	0.60	0.60
2.	Visit	1	99.60	99.60
3.	Interaction, Change by Visit	1	6.00	6.00
4.	Subject	6	455.32	75.89
5.	Interaction, Change by Subject	6	365.56	60.93
6.	Interaction, Visit by Subject	6	209.10	34.85
7.	Interaction, Change by Visit by Subject	6	257.60	42.93
8.	Teacher	58	8,849.60	152.58
9.	Change by Teacher	58	1,812.19	31.24
10.	Interaction, Visit by Teacher	58	1,673.87	28.86
11.	Residual	58	1,831.90	31.58
12.	Sub-total	259	15,561.34	
13.	Pooled Residual	116	3,486.06	30.05

$$r_{xx} = \frac{152.58 - 30.05}{152.58} = .80$$

Green and Tukey, 1960). In Table 15, this pooled residual mean square, 30.05, was used in calculating the estimate r_{xx} of the stability of scores on the key.

This coefficient may be interpreted as an estimate of the correlation between a score based on four visits (of approximately 28 minutes' duration) to a teacher and the score that would be obtained on four other visits, each made within a week or so of one of the first four visits, by the same observer.

Table 16 shows the estimated stability coefficients of all 41

Insert Table 16 about here

---

contrasts, together with other descriptive data. The median coefficient is .58, and individual values range all the way from zero to .85.

The hypothesis that a key does not detect any stable differences between teachers when four 28-minute visits are made) was tested by comparing the mean squares for Teachers and either the Teacher by Visit interaction or Residual mean square; it was rejected at the .05 level for all but four of the 41 contrasts.

For comparison, raw frequencies on the 42 original categories were analyzed in the same way, with the results shown in Table 17.

Insert Table 17 about here

The median reliability here is .50, and the hypotheses of no

Table 16

Descriptive Data on 41 OScAR Contrasts

	Contrast Name	Stability Coefficient	Mean and Range of Scores Per Visit: Mean	Lowest Score	Highest Score	Range
1.	Affective Statements	.70**	+9.9	0	+47	47
2.	Positive Affect	.80**	+0.3	-43	+25	68
3.	Concern	.70**	-4.2	-25	+1	26
4.	Hostility	.82**	-2.9	-26	+6	32
5.	Pupil Talk	.77**	+5.4	0	+227	227
6.	Continuing Pupil Statements	.78**	+1.6	-26	+223	249
7.	Management	.32*	+21.2	0	+126	126
8.	Control	.34*	+12.0	-56	+126	182
9.	Dominance	.45**	+2.5	-91	+47	138
10.	Direction Complexity	.31*	-4.4	-45	+16	61
11.	Description Complexity	.20	-0.8	-17	+76	93
12.	Permissiveness	.44**	+1.5	-81	+24	105
13.	Pupil Initiative (Procedural)	.35*	+1.2	-18	+81	99
14.	Benevolent Despotism	.00	+1.1	-15	+81	96
15.	Teacher Talk	.77**	+26.1	-118	+251	369
16.	Suspense	.85**	-9.5	-270	+81	351
17.	Problem Complexity	.50**	-15.2	-61	+56	117
18.	Lecturing	.80**	8.5	-30	+236	266
19.	Teacher Initiative (Substantive)	.78**	+5.1	-47	+51	98
20.	Cohesion	.66**	-3.5	-48	+61	109
21.	Divergence	.58**	-21.0	-123	+43	166
22.	Feedback	.68**	-10.4	-77	+59	136
23.	Valence	.78**	+15.9	-33	+140	173
24.	Enthusiasm	.31*	-16.7	-122	+19	141
25.	Encouragement	.62**	-19.0	-63	+17	80
26.	Positivity	.50**	+2.2	-120	+67	187
27.	Feedback by Teacher Initiative	.68**	+0.6	-23	+33	56
28.	Feedback by Cohesion	.35*	+0.1	-37	+31	68
29.	Feedback by Divergence	.61**	+3.2	-60	+67	127
30.	Valence by Teacher Initiative	.59**	+3.1	-16	+45	61
31.	Valence by Cohesion	.05	-1.1	-59	+55	114
32.	Valence by Divergence	.57**	-6.7	-120	+12	132
33.	Enthusiasm by Teacher Initiative	.40*	-2.5	-39	+13	52
34.	Enthusiasm by Cohesion	.40*	-5.5	-61	+13	74
35.	Enthusiasm by Divergence	.37*	+7.8	-15	+110	125
36.	Encouragement by Teacher Initiative	.72**	-0.1	-22	+37	59
37.	Encouragement by Cohesion	.53**	+1.4	-20	+23	43
38.	Encouragement by Divergence	.63**	+7.4	-18	+53	71
39.	Positivity by Teacher Initiative	.60**	-2.6	-40	+27	67
40.	Positivity by Cohesion	.32	+0.7	-23	+55	78
41.	Positivity by Divergence	.31*	+0.3	-50	+109	159

*$.01 < P < .05$

**$P < .01$

Table 17

Descriptive Data on 41 OScAR Categories
Shown in Order of Stability

Category	Stability Coefficient	Rate Per Hour Mean	Rate Per Hour Range
Rebuking, Initiating	.84	7.7	0-41
Informing, Continuing	.82	51.7	3-325
Pupil Statement, Continuing	.78	6.9	0-179
Elaborating, Approved	.76	7.8	0-43
Substantive Pupil Init., Acknowledge	.76	8.9	0-62
Convergent, Acknowledged	.75	17.5	0-88
Convergent, Support	.70	3.5	0-24
Informing, Initiating	.68	34.8	12-61
Problem Structuring, Cont.	.68	18.5	1-101
Considering, Initiating	.67	9.3	0-32
Convergent, Not Evaluated	.66	5.5	0-20
Divergent, Approved	.65	3.6	0-17
Problem Structuring, Init.	.65	48.9	10-105
Substantive, Pupil Init., Approved	.63	4.7	0-21
Elaborating, Supported	.61	2.3	0-36
Elaborating, Acknowledged	.56	8.1	0-29
Convergent, Criticized	.55	1.4	0-16
Substantive Pupil Init., Neutrally Rej.	.55	1.5	0-9
Describing, Initiating	.53	7.8	1-34
Elaborating, Neutrally Rejected	.51	1.0	0-9
Directing, Initiating	.50	14.0	0-33
Rebuking, Continuing	.49	1.5	0-13
Convergent, Neutrally Rejected	.48	4.6	0-16
Substantive Pupil Init., Not Evaluated	.48	1.1	0-8
Substantive Pupil Init., Supported	.48	.9	0-10
Pupil, Initiating	.47	3.8	0-18
Divergent, Criticized	.46	.1	0-2
Convergent, Approved	.45	19.3	1-78
Divergent, Not Evaluated	.36	1.1	0-13
Describing, Continuing	.35	6.2	0-66
Non-Substantive, Pupil Initiated, Posit.	.27	3.1	0-13
Divergent, Acknowledged	.23	3.1	0-13
Substantive, Pupil Initiated, Criticized	.19	.7	0-14
Non-substantive, Pupil Init., Negative	.16	2.7	0-43
Divergent, Neutrally Rejected	.15	.7	0-6
Directing, Initiating	.15	14.0	0-33
Elaborating, Not Evaluated	.14	2.8	0-13
Non-Substantive Teacher Init., Posit.	.13	3.0	0-21
Divergent, Supported	.08	.7	0-6
Non-substantive Teacher Init., Negative	.05	.4	0-3
Elaborating, Criticized	.00	.6	0-6
Considering, Continuing	.00	.8	0-6

reliability was accepted for 12 of the categories. The range of reliabilities is from zero to .84.

It should be noted that on the average the contrasts are appreciably more stable than the categories upon which they are based! This is probably attributable to the fact that although, if an observer misclassifies a behavior, the frequencies of both categories change, some contrast scores may be unaffected by the error.

Misclassifying an Approving exit as Supporting, for instance, will not affect scores on Key 22, Feedback or on Key 23, Valence. The stability of Feedback or Valence scores is, then, not affected by confusion between Approving and Supporting exits (or between Divergent and Convergent entries, for that matter). Apparently the characteristics measured by contrasts are naturally more internally consistent and hence more stable than the category frequencies recognized by the observers!

Inspection of the mean scores and ranges on the categories measured least reliably, as shown in Table 17, suggest that in several cases failure to obtain a stable estimate of the frequency of behaviors in a category was probably due to the fact that such behaviors were rarely observed. Behaviors in each of the six most unstable categories occurred no oftener than three times per hour--i.e., once every 20 minutes. This means that the behavior might be expected to occur only once per visit in the average class. The range of frequencies of unreliable categories also tends to be narrow, running from zero to no more than six per hour in four

instances.

Behaviors in the six most stable categories, on the other hand, tended to occur at least twice as often, and their ranges were also at least twice as great.

There are exceptions, however. Among categories with low reliabilities is Directing, Initiating, with a range from zero to 33, a mean of 14, and a reliability of only .15; among the more stable ones is Neutrally Rejected Pupil-Initiated Substantive, with a range from zero to nine, a mean of 1.5 per hour, and a reliability of .55. The least common item of all, Divergent, Criticized, with a mean frequency of once in ten hours, and a range from zero to only twice an hour, had a stability coefficient of .46, significant at the .01 level. Stability of the behavior--i.e., consistency in the behavior of a teacher from week to week--may be high even when the average frequency is low.

Preliminary Analysis of FIAT Data

Over the years, users of FIAT have developed a large number of scores called measures, most of which are based on combinations of two or more matrix cells, formed on a priori grounds. From a list of such measures compiled by Soar (personal communication) a number were chosen for use in this study. In addition, some new measures were devised which seemed to correspond to items in OScAR. A total of 44 measures were scored on FIAT and used in the preliminary analysis. Table 18 defines each measure by indicating

---

Insert Table 18 about here

Table 18

Compositions and Stability Coefficients of Variables

Scored on FIAT Records

Scale No.	Variables	Name	Stability Coefficient
1.	Accepts Feeling	Total in Category 1	.40*
2.	Praises	Total in Category 2	.80**
3.	Accepts Ideas	Total in Category 3	.80**
4.	Asks Questions	Total in Category 4	.80**
5.	Lecturing	Total in Category 5	.74**
6.	Directing	Total in Category 6	.66**
7.	Criticizing	Total in Category 7	.71**
8.	Student Responds	Total in Category 8	.83**
9.	Student Initiates	Total in Category 9	.80**
10.	Silence or Confusion	Total in Category 10	.70**
11.	Indirect Teacher Talk	Cells 11, 12, 13, 21, 22, 23, 31, 32, 33	.71**
12.	Direct Teacher Talk	Cells 66, 67, 76, 77	.23
13.	Indirect vs. Direct Influence	Variable 11 - Variable 12	.71**
14.	Extended Indirect vs. Direct Influence	Variables 1, 2, 3 - Variables 6,7	.65**
15.	Content Cross — Content	Variables 4, 5	.80**
16.	Center of Cross	Cells 44, 45, 54, 55	.73**
17.	Extended Praise	Cell 22	.08
18.	Extended Criticism	Cell 77	.51**
19.	Extended Questioning	Cell 44	.68**
20.	Extended Lecturing	Cell 55	.73**
21.	Extended Directing	Cell 66	.15
22.	Elaborating Questions	Cell 34	.84**
23.	Difficult Questions	Cell 4-10	.42**
24.	Pacing	Cell 5-10	.46**
25.	Sustained Student Talk	Cells 88, 89, 98, 99	.73**
26.	Student Response Praised	Cell 82	.77**
27.	Student Response Accepted	Cells 81, 83	.85**
28.	Student Response Not Evaluated	Cells 84, 85, 86	.72**
29.	Student Response Criticized	Cell 87	.66**
30.	Student Initiation Praised	Cell 92	.61**
31.	Student Initiation Accepted	Cells 91, 93	.67**
32.	Student Initiation Not Evaluated	Cells 94, 95, 96	.56**
33.	Student Initiation Criticized	Cell 97	.07
34.	Convergent Question	Cell 48	.84**
35.	Divergent Question	Cell 49	.63**
36.	Directed Student Response	Cell 68	.45**
37.	Directed Student Initiation	Cell 69	.22
38.	Student Response After Indirect Talk	Cells 18, 28, 38	.43**
39.	Student Initiation After Indirect Talk	Cells 19, 29, 39	.66**
40.	Student Response After Lecturing	Cell 58	.25
41.	Student Initiation Following Lecturing	Cell 59	.68**
42.	Silence or Confusion After Student Initiation	Cells 8-10, 9-10	.63**
43.	Indirect vs. Direct Following Student Response	Cells 81, 82, 83 - Cells 86, 87	.85**
44.	Indirect vs. Direct Following Student Initiation	Cells 91, 92, 93 - Cells 96, 97	.66**

*$.01 < P < .05$

**$P < .01$

which cells enter into it and also gives its stability coefficient. Again the problem of lack of experimental independence appears, since there is considerable overlap in items on different measures, substantial (and sometimes high) spurious correlations between measures may be expected.

Certain of the measures listed by Soar were based on ratios on the form a over b. Such a ratio is usually referred to as an "a/b ratio." Because ratios of numbers, both of which are subject to error, have undesirable statistical properties, each ratio of this form was replaced in the present study by a contrast of the form a - b. This was a second major departure from normal procedures for using FIAT data; it should be borne in mind in interpreting the results of the study and relating them to other research using FIAT.

The data collected in this study with FIAT included 280 complete records on 70 teachers, and each FIAT record was based on the same behavior sample as one of the 280 OScAR records. Each FIAT record was scored on the 44 "measures" described in Table 18, and these 44 x 280 = 12,320 scores constituted the FIAT data.

The 280 scores on each measure were analyzed in the same design as was used with the OScAR data, and a stability coefficient was calculated for each measure, with the results shown in Table 18. The median stability coefficient of the 44 measures was .68, with a range from .07 to .85; and all but six of the coefficients differed significantly from zero. It should be noted that the stability coefficients of the ten basic categories used by the FIAT

coder (which appear in Table 18 as variables one to ten, inclusive) have stabilities ranging .40 to .80 (median about .79(.

Factor Analysis

All 85 variables--41 OScAR contrasts and 44 FIAT measures--were intercorrelated across the 280 observations made of the 70 teachers, and the matrix was submitted to a principal components analysis (Hotelling, 1933). Communalities were estimated by Tucker's method using the adjusted highest off-diagonal element, and ten components were extracted with eigen-values greater than one. The ten components were rotated by the varimax method to orthogonal simple structure. Table 19A shows the rotated loadings on 32 OScAR contrasts and TAble 19B shows loading on 22 FIAT measures. Variables (either on OScAR or FIAT) on which there was no loading whose absolute value was greater than or equal to .50 are not shown. Variables in each table have been grouped so the

---

Insert Tables 19A and 19B about here

---

those which load on the same factor are next to each other.

The principal purpose of this factor analyses was to reduce the number of variables or scores to be submitted to further analyses. The next step in accomplishing this purpose was to develop for each instrument a simplified key designed to approximate a factor score for each factor that loaded on that instrument.

Factor 1, for example, had substantial loadings on both instruments, so a key was constructed for each. In constructing the

Table 19A

Rotated Factor Loadings on 32 OScAR Keys (Decimal Points Omitted)

No.	OScAR Contrast	Factor I	II	III	IV	V	VI	VII	VIII	IX	X
16.	Suspence	-75*	-23	00	07	00	-13	-14	-25	-07	09
15.	Teacher Talk	72*	16	-07	-10	-10	-05	14	-06	-02	-17
18.	Lecturing	63*	19	01	-10	-05	02	10	01	04	-15
38.	Encouragement by Divergence	-07	-81*	-08	14	-01	-02	-01	-07	12	-05
21.	Divergence	13	73*	-24	23	-09	17	05	06	-01	-02
25.	Encouragement	07	67*	16	-10	03	11	-17	-24	00	07
20.	Cohesion	-06	53*	-10	02	-19	-04	02	21	-14	-11
37.	Encouragement by Cohesion	-01	-58*	00	-07	06	12	02	-20	06	02
35.	Enthusiasm by Divergence	-01	-53*	70*	-13	10	-03	-03	00	06	02
41.	Positivity by Divergence	00	10	87*	01	06	04	-02	08	-06	01
26.	Positivity	-00	-18	-82*	03	-01	-01	10	17	02	06
29.	Feedback by Divergence	-01	-36	-80*	-05	-04	-04	-08	-09	-05	-01
32.	Valence by Divergence	09	37	-79*	13	09	03	-04	-05	-04	-04
22.	Feedback	04	28	73*	-22	01	-02	02	-14	13	03
23.	Valence	-17	-10	69*	-06	-29	-15	15	04	-01	04
24.	Enthusiasm	04	37	-69*	-03	03	13	-07	-03	-06	02
30.	Valence by Teacher Initiative	-19	-34	57*	-08	-14	-07	08	-04	-13	-09
33.	Enthusiasm by Teacher Initiative	06	33	-52*	09	01	05	-03	19	03	06
28.	Feedback by Cohesion	-03	-15	-51*	-12	-12	-05	-01	-13	06	08
4.	Hostility	03	-02	05	07	-68*	-01	-02	-16	11	-06
1.	Affective Statements	-01	07	-06	-08	65*	-01	-04	22	-14	31
2.	Positive Affect	-04	07	02	01	-58*	05	01	-07	15	37
5.	Pupil Talk	-22	11	01	-08	-10	85*	-07	10	-12	-05
6.	Continuing Pupil Statements	-13	05	03	-07	-12	77*	-03	07	-14	-09
19.	Teacher Initiative (Substantive)	-26	-34	23	-18	-22	-20	-01	-66*	-04	01
36.	Encouragement by Teacher Initiative	13	40	-01	14	18	15	00	66*	-10	-05
27.	Feedback by Teacher Initiative	12	13	41	01	03	19	12	63*	11	05
39.	Positivity by Teacher Initiative	-07	-05	-48	-03	-11	-19	-06	-52*	02	02
13.	Pupil Initiative (Procedural)	05	07	02	05	10	11	-01	-01	-84*	-06
12.	Permissiveness	-02	-04	02	-04	-14	-11	-01	09	77*	01
7.	Management	08	03	-03	-10	35	06	-08	-01	-28	65*
8.	Control	01	-11	-07	-14	19	-02	00	-05	23	61*
3.	Concern	09	-13	02	05	01	02	02	-11	-04	-57*

* Loading with absolute value equal to .50 or greater.

Table 19B

Rotated Factor Loadings on 22 FIAT Measures

FIAT Measure	Factor I	II	III	IV	V	VI	VII	VIII	IX	X
Extended Lecturing	.92*	-.01	-.03	-.08	-.18	.05	-.20	.10	-.04	.05
Lecturing	.88*	.06	-.03	-.08	+.13	.13	-.25	.18	.01	.07
Censor of Content Cross	.85*	.02	-.02	-.09	-.24	.18	-.18	-.01	.07	.04
Student Responds	-.63*	.19	.02	-.46	-.13	-.06	.05	.03	00	-.10
Asks Question	-.55*	.37	.03	-.13	-.15	.55*	.08	-.23	-.11	-.07
Content Cross	-.47	.49	.03	-.15	.03	.56*	-.07	.10	00	-.02
Student Response Accepted	-.15	.66*	-.17	-.15	-.10	.26	.15	-.08	-.07	-.11
Elaborating Question	-.28	.58*	-.04	-.06	-.14	.24	.31	-.12	-.12	-.08
Indirect vs. Direct After Student Response	-.27	.56*	.01	-.28	-.28	.39	.33	-.03	-.07	-.11
Accepts Ideas	-.10	.51*	-.05	.16	-.10	.20	.77*	.01	-.06	-.06
Indirect Teacher Talk	-.03	.10	.10	.12	-.09	.04	.86*	-.03	.01	.02
Indirect vs. Direct Influence	-.00	.08	.08	.16	-.31	.03	.84*	.03	-.02	-.08
Student Initiates	-.15	-.20	-.03	.84*	-.12	-.20	-.15	.15	-.08	-.09
Student Initiation Praised	-.06	-.15	.10	.65*	-.18	-.03	-.15	.02	.02	.01
Student Initiation Accepted	00	.07	-.14	.66*	-.05	.03	.21	.33	-.05	-.03
Divergent Question	-.13	-.13	-.07	.56*	-.16	.04	-.13	-.07	-.12	-.03
Student Initiation After Indirect Talk	-.03	-.08	00	.56*	-.15	00	.21	.31	00	-.12
Criticizing	-.07	.06	00	-.06	.89*	.10	-.06	.10	-.04	00
Extended Indirect vs Direct Influence	-.03	.18	-.02	.34	-.72*	.17	.36	.13	-.07	-.21
Extended Criticism	-.03	00	.05	-.14	.61*	.02	-.11	-.06	-.02	-.02
Silence or Confusion	-.10	-.15	-.02	-.24	.55*	-.11	-.16	-.03	.16	.11
Student Initiation After Lecturing	.17	.07	-.05	.34	.06	.03	-.27	.60*	.04	00

*Loading with absolute value equal to .50 or greater

keys, only loadings with absolute magnitudes of .50 or more were used. This value was chosen as a cutoff point mainly because it seemed to represent a natural gap, there being relatively few loadings between .40 and .50 anywhere in Table 19.

Five FIAT measures met this criterion on Factor I: 20. Extended Lecturing, 5. Lecturing, and 16. Center of Content Cross, with positive loadings; 8. Student Response, and 4. Asks Questions, with negative ones. Study of Table 18 (above) suggests that measures based primarily on behaviors classified in Flanders' Category, 5. Lecturing, show high positive loadings while variables based on Categories 4. Asks Questions and 8. Student Talk--Response show substantial negative loadings.

As a scoring key for this factor, called Lecturing, the difference between the frequency in Cell 55 of the interaction analysis matrix and the total in Category 8. Student Talk--Response was adopted. High-scoring teachers on this dimension would be ones who frequently talked for six seconds in a row, and whose pupils answered teacher questions rather seldom.

Factor I also had three substantial loadings on OScAR; a negative loading on 16. Suspense, and positive loadings on 15. Teacher Talk and 18. Lecturing.

The procedure used in devising factor keys on OScAR was to add the weights used on all contrasts that loaded .50 or more on the factor (with indicated signs) to get a new set of weights, identifying a new contrast. This "factor contrast" is, of course, orthogonal to all contrasts that are orthogonal to those from

which it was derived (and therefore to all of the other factor contrasts). Table 20 illustrates how this was done, using Factor I as an example.

Insert Table 20 about here

Weights used in the resulting contrast, named Lecturing Behavior, may then be inspected for meaning. In this instance, is apparent that scores on this key mainly reflect Continuing Informing Statements and that substantive interchanges (in which students normally answer teacher questions; reduce a teacher's score.

On the surface, the two keys derived from this factor--one for each instrument--resemble one another rather closely (as, indeed, they should).

Factor II loads positively on four FIAT measures: 27. Student Response Accepted, 22. Elaborating Question, 43. Indirect vs. Direct Following Student Response, and 3. Accepts Ideas. From Table 18 it appears that 3. Accepts Ideas when preceded by 8. Student Talk--Response or followed by 4. Teacher Question contains the essence of this dimension (3. Accepts Ideas without 8 or 4 has a loading of only .51); and that 86 and 87 have a negative relationship to it. Accordingly, the difference between the total frequencies in Cells 81, 82, 83, and 34 and the total in Cells 86 and 87 of the matrix was adopted as the second FIAT factor key, and called (because it is identical to FIAT measure number 43) the

Table 20

Derivation of A Contrast Key To Approximate A Factor Score

Contrast Key	Category				
	Informing Statements		Problem-Structuring Statements		Substantive Interchanges
	Initiating	Continuing	Initiating	Continuing	
15. Teacher Talk	+1	+1	+1	+1	-1
16. Suspense (reflected)	+1	+1	-1	-1	0
18. Lecturing	-1	+1	0	0	0
Lecturing (Total)	+1	+3	0	0	-1

I-D Contrast on Student Response. (This and all of the other factor keys scored on FIAT are shown in Table 21.)

Insert Table 21 about here

Factor II also had interpretable loadings on six OScAR contrasts (see Table 19A). When the weights on these contrasts were combined (and the totals simplified slightly) a factor key called Question Quality was obtained. (This contrast, together with the other seven scored on OScAR, is shown in Table 22.) This key bears little superficial resemblance to the Factor II key just

---

Insert Table 22 about here

---

derived for FIAT. For one thing, it is much more complex--involving, as it does, all 24 categories of substantive interchanges, with weights from -16 to +4.

Lowest weights are assigned to Convergent Interchanges Accepted, Approved, or Criticized. Highest weights are assigned to interchanges either entered with an Elaborating Question or exited from with No Evaluation, that is, followed immediately by another question.

Typical behavior of a teacher high on this key, then, is to react to a pupil response by asking another question instead of evaluating the response. Rather than tell a student he is wrong (or right, as the case may be) this teacher either asks a probing

Table 21

Summary of Factor Keys Scored on FIAT

Factor Number	Name	Composition
I	Lecturing Behavior	(Cell 55)-(Category 8 total)
II	I-D Contrast on Student Response	(Cells 81, 82, 83, 34) minus (cells 86, 87)
IV	Student Initiations	Category 9 total
V	Criticizing Behavior	Category 7 total
VI	Modified Content Cross	(Twice Category 4 total) + (Twice Cells 44, 45, 54, 55) + (Category 5 total)
VII	Indirect Teacher Talk	Cells 11, 12, 13, 21, 22, 23, 31, 32, 33
VIII	Student Interruptions	Cell 59

Table 22

Scoring Weights for Eight Factor Contrasts
Scored on OScAR 4V

Factor I
Lecturing Behavior

	Informing Statements Initiating	Statements Continuing	Substantive Interchange
Weights	+1	+3	-1

Factor II
Question Quality

	Supported	Approved	Accepted	Not Evaluated	Neutrally Rejected	Criticized	Total
Convergent	-1	-11	-16	+4	-1	-11	-36
Divergent	+3	+1	0	+4	+3	+1	+12
Elaborating	+4	+4	+4	+4	+4	+4	+24
Pupil Initiated	+2	-2	-4	+4	+2	-2	0
Totals	+8	-8	-16	+16	+8	-8	0

Factor III
Question Difficulty

	Supported	Approved	Accepted	Not Evaluated	Neutrally Rejected	Criticized	Total
Convergent	-9	+95	-54	-6	-9	-17	0
Divergent	+3	+11	-6	-6	+3	-5	0
Elaborating	-12	+44	-12	+12	-12	-20	0
Pupil Initiated	-6	+18	-24	0	-6	+18	0
Totals	-24	+168	-96	0	-24	-24	0

Factor V
Rebuking Behavior

	Rebuking Statements Initiating	Continuing
	+3	+1

Table 22 (continued)

Factor VI
Listening Behavior

Pupil Statements
Continuing

+1

Factor VIII
Question Source

	Supported	Approved	Accepted	Not Evaluated	Neutrally Rejected	Criticized	Totals
Convergent	-1	0	- 4	0	-1	0	- 6
Divergent	-1	0	- 4	0	-1	0	- 6
Elaborating	-1	0	- 4	0	-1	0	- 6
Pupil Initiated	+3	0	+12	0	+3	0	+18
	0	0	0	0	0	0	0

Factor IX
Permissive Behavior

Non-Substantive Interchanges				
Teacher-Initiated		Pupil-Initiated		
Positive	Negative	Positive	Negative	Total
+2	0	0	-2	0

Factor X
Managing Behavior

	Describing	Directing	Considering	Totals
Initiating	+2	+2	+1	+5
Continuing	+2	+2	-1	+3
Totals	+4	+4	0	+8

question of the same student or turns to another student and asks him to comment.

The low-scoring teacher, on the other hand, tends to ask questions that are more discrete, less connected: usually he acknowledges a pupil's answer before going on to ask another question; at other times he gives feedback, critical when the student's answer is incorrect, and rarely supportive.

The key is bipolar, contrasting two distinct questioning styles--one rapid-fire, disconnected; the other more deliberate, connected, and likely to call for more thoughtful pupil answers.

Factor III appears only on OScAR, where it loads no fewer than 11 contrasts. The composite, called Question Difficulty, as shown in Table 22, like the last one described, is based on all 24 substantive interchanges, and is complex in nature.

The basic contrast appears to be between either Convergent or Elaborating Interchanges, Approved and Convergent Interchanges, Accepted. It should also be noted that Approved exits receive a total of 168 plus points, while all other exits (except ones Not Evaluated) are predominantly negatively weighted--especially Accepted ones. To get a high score on this dimension a teacher should ask easy questions, that is, ones which will elicit student responses he can approve but which do not merit support or praise (as they would if the questions were difficult ones). Hence, the name given to the key.

Factor IV loads only on FIAT, where five variables show interpretable loadings: 9. Student Initiates, 30. Student Initiation

Praised, 36. Student Initiation Accepted, 35. Divergent Question, and 39. Student Initiation After Indirect Talk. Category 9. Student Talk--Initiation seems to be the one element common to all five measures, so this key was defined simply as the total frequency in Category 9, and called Student Initiations.

Factor V loaded on four FIAT variables and three OScAR contrasts. The highest FIAT loading was a positive one on Measure 7. Criticizing; the factor key was called Criticizing Behavior, accordingly, and defined as the total recorded in Category 7.

On OScAR, negative loadings were found on both 4. Hostility and 2. Positive Affect, plus a positive one on 1. Affective Statements. On their face, these loadings seem somewhat contradictory; but inspection of the weights in the Factor Contrast obtained by combining them, (+3 on Initiating and +1 on Continuing Rebuking Statements) clears up this seeming contradiction. The dimension measured identifies teachers who rebuke pupils frequently but with little hostility or rancor. This key was named Rebuking Behavior.

Factor VI also appears on both instruments. On FIAT it shows interpretable, though weak, loadings on 4. Asks Questions and 15. Content Cross. Since no strong impression comes through as to the nature of this factor, the two measures were combined and named Modified Content Cross. (From the much larger OScAR loading reported below, this factor seems to have to do primarily with pupil answers to teacher questions.)

On OScAR, the factor contrast turns out to be the number of

Continuing Student Statements the teacher permits. In other words, the longer a teacher permits a pupil to talk before interrupting him, the higher his score on this factor key. The key was named Listening Behavior.

Factor VII, like Factor IV, did not appear on OScAR. On FIAT it showed a high positive loadings on three measures: 3. Accepts Ideas, 11. Indirect Teacher Talk, and 13. Indirect vs. Direct Influence. All three of these overlap in the upper left-hand corner of the interaction analysis matrix, which suggest that these cells account for the loadings. The key was accordingly assigned the name already given to it by Flanders, Indirect Teacher Talk (cf. Flanders, 1965).

Factor VIII loads on only one cell in the matrix, the one referred to as Measure 41. Student Initiation After Lecturing--Cell 59. As a factor key, it was named Student Interruptions.

On OScAR this factor loads on four contrasts, which when combined yield the one shown in Table 22 as Question Source. Notice that this key is a contrast based on 12 kinds of interchanges. The meaning of this variable, defined on FIAT simply in terms of one student behavior--making a non-responsive comment while the teacher is talking--is considerably illuminated by the weights on OScAR, where there is evidence about how the teacher reacts to such initiations, and also evidence that the teacher whose students often interrupt him exhibits not only a characteristic way of reacting to the interruptions but also a particular pattern in evaluating student responses to his own questions. Even though a cause-effect

relationship has not been demonstrated, these weights do suggest how a teacher might go about encouraging--or discouraging--such behavior.

Factor IX appears only on OScAR, where it loads on two contrasts. From Table 22 it appears that this dimension contrasts teachers who frequently offer pupils a choice of action and rarely refuse permission when asked--"permissive" teachers--with teachers who rarely offer pupils a choice of action and frequently refuse permission when asked for it. It has been named Permissive Behavior.

Factor X, which is also unique to OScAR, loads on three contrasts, and yields the composite shown in Table 22. The teacher high on Factor X spends much time in talking about procedure, sometimes using statements that although they are coded as Considering do not seem to reflect any genuine positive affect. They are probably requests for pupils to do something which the pupils react to as Directing Statements, but which the observer codes as Considering. It has been named Managing Behavior.

Discussion

The analysis just described had two primary purposes. One was to reduce the number of scoring keys needed to summarize the information in an OScAR record. The other was to learn something about the nature of the variables measured by the instrument by comparing it with a better known observation schedule, Flanders' Interaction Analysis Technique, or FIAT.

With respect to the latter point it seems fair to say that the

two instruments measure similar aspects of classroom behavior. From Table 23, which shows the percent of variance in the matrix

---

Insert Table 23 about here

---

of intercorrelations that is accounted for by each factor, it appears that in the 10 factor scores about 60% of the variance in both instruments is accounted for. If this amount of variance--the "explained" variance, is used as a base, then it may be said that factors common to both instruments (I, II, IV, VII, and VIII) make up 56% of the variance in both systems. Since the factors are orthogonal, this means that the remaining 44% may be divided into two portions which are unique to one or the other instrument. The factors unique to OScAR (III, IX, and X) comprise 24%, and the factors unique to FIAT (IV and VII) account for 20%. To put it another way, about 70% of the variance in dimensions measured by either instrument is also found on the other.

In interpreting these figures, it should be borne in mind that they are based on a factor analysis which sought to maximize common variance. Thus the figure of 56% is an upper limit on overlap between the two systems. It would not be far from the truth, perhaps, to say that in this population no more than 2/3 of the systematic variance of each instrument is shared by the other, while at least a third is unique.

Variance detected by FIAT but not by OScAR includes the variable of teacher directness or indirectness which is so important

Table 23

Percent of Variance Accounted for by each of 10 Factors

Factor	Root	Percent of Total Variance	Percent of "Explained" Variance
I	6.78	9.0	15.1
II	6.59	8.8	14.8
III	6.48	8.6	14.4
IV	5.09	6.8	11.4
V	4.93	6.6	11.1
VI	3.95	5.3	8.9
VII	3.68	4.9	8.2
VIII	3.07	4.1	6.9
IX	2.07	2.8	4.7
X	2.01	2.7	4.5
	TOTAL "Explained" variance	59.6	

in interaction analysis theory. While OScAR does seem to be sensitive to the degree to which a teacher is accepting and pupil-supportive as a function of the type of pupil behavior he is reacting to, OScAR does not detect the tendency to accept whatever pupils do which appears as Factor VII.

More easy to explain is the failure of OScAR to detect the difference between Flanders' categories 8 and 9 which is implied in Factor IV, since OScAR 4V coders do not record pupil responses at all.

The fact that the Flanders' records do not reflect Factor III, called Question Difficulty, is also somewhat unexpected. Perhaps this fact and the insensitivity of OScAR to indirect teacher talk both reflect a difference in the way the two systems look at teacher reactions to student behavior. Flanders' coders seem primarily to be sensitive to what might be called psychological acceptance or support. OScAR coders are expected to discriminate cognitive from affective reactions--feedback about factual accuracy from either praise or critical rejection.

Factors IX and X probably elude FIAT because they relate to procedural matters and are affect free.

The 15 scoring keys developed from the factor loadings include five pairs of keys based on the same five factors. However, the process used in deriving them, and the nature of the contrasts or measures which load on each factor, leads in some instances to pairs of keys which bear little superficial resemblance to one another despite their common ancestry. Table 24 shows the inter-

Insert Table 24 about here

correlations within and between these pairs of keys.

Because of their common origin, one would expect the keys to show high correlations on the diagonals. The keys based on Factors I and V follow this pattern, correlating high within pairs and low outside of them, but the rest do not. The Modified Content Cross key (scored on FIAT) correlates substantially higher with three other OScAR keys than with the one derived from the same factor. Inspection of Table 25 indicates that it also correlates strongly

Insert Table 25 about here

with the FIAT keys based on Factors I, Lecturing Behaviors, and II, I-D Contrast on Student Response. From Table 21 it appears that the heavy weights attached by both the key based on Factor VI and the one based on Factor I to category 5 is probably responsible for the correlation between the two. The correlation between the FIAT keys based on VI and II may also be an artifact resulting from the strong tendency for 4. Asks Questions to be followed by 8. Student Talk--Response.

Similarly, the correlations between FIAT key VI and Question Source and Question Quality (both scored on OScAR) probably reflect the fact that all of them are based on teacher questions. Apparently the approximation to Factor VI provided by the Modified

Table 24

Correlations Between Keys Based on Factors Common to OScAR and FIAT

OScAR Keys	FIAT KEYS I Lecturing Behavior	II I-D Contrast on Student Response*	V Criticizing Behavior	VI Modified Content Cross	VIII Student Interruptions
I. Lecturing Behavior	64	-32	-19	46	16
II. Question Quality	27	-44	-17	52	06
V. Rebuking Behavior	-10	-12	77	-01	12
VI. Listening Behavior	-18	-17	-08	23	-03
VIII. Question Source	27	-48	23	45	38

*Reflected

Table 25

Intercorrelations Between Factor-Based Keys Scored on FIAT

	I	II	IV	V	VI	VII	VIII
I. Lecturing Behavior		-44	11	-17	59	07	26
II. I-D Contrast on Student Response			-34	-10	-73	20	-18
IV. Student Initiations				-12	24	-04	37
V. Criticizing Behavior					-13	-26	08
VI. Modified Content Cross						08	12
VII. Indirect Teacher Talk							-19
VIII. Student Interruptions							

Content Cross key is a poor one.

The high negative correlation between the I-D Contrast on Student Response and Question Source, which is larger than the correlation between the former key and Question Quality (derived from the same factor), suggest that the more indirect (i.e., supporting) the teacher is toward student response, the fewer questions students ask. This sounds as though the teacher may be extinguishing pupil questions by rewarding pupil responses to teacher questions.

The negative correlation (-.44) between the I-D Contrast on Student Response and Question Quality, which are supposed to be based on the same factor indicates that the "factor key" for FIAT does not capture the meaning of that factor very successfully. For what it is worth, this correlation suggests the teacher who is indirect in responding to pupil questions may be asking many convergent questions and reinforcing pupils' responses to them; while the "direct" teacher may be asking questions that call for more thought (mostly elaborating questions) and withholding approval to stimulate such thought on the part of pupils.

A more careful inspection of Tables 25 and 26 (which show

Insert Table 26 about here

---

intercorrelations between keys based on different factors and scored on the same instrument) reveals a generally satisfactory pattern of low correlations, except for the effects of the Modified

Table 26

Intercorrelations Between Factor-Based

Keys Scored on OScAR

		I	II	III	VIII	V	VI	IX	X
I.	Lecturing Behavior		20	-02	24	-11	-07	-02	-16
II.	Questioning Style			-08	26	-08	14	-10	-11
III.	Question Difficulty				08	-02	-07	05	-15
VIII.	Direction Source					28	13	-03	-04
V.	Rebuking Behavior						-05	-14	24
VI.	Listening Behavior							-12	06
IX.	Permissive Behavior								00
X.	Managing Behavior								

Content Cross already noted. The correlation of -.44 between Lecturing and the I-D Contrast on Student Response shown in Table 25 is doubtless an artifact of the overlap between the two keys. It is difficult to avoid such overlap when there are only ten basic categories in the first place!

To sum up: as far as can be told from limited data, all of the keys seem potentially useful except one--the Modified Content Cross; and despite their common origin, all of the 14 remaining keys seem to describe rather different aspects of classroom behavior, except that the two Lecturing keys and the two Rebuking and Criticizing keys each yield rather similar information, though scored on different instruments. These results should be cross-validated before being regarded as anything more than tentative; however, most of the keys reported have sufficient intrinsic interest to merit further testing on their own merits.

The fact that the keys developed for Flanders' system (especially I, II, IV, V, and VII) bear so close a resemblance to ones recommended on a priori grounds by the originator of that system is prima facie evidence of their validity as well as of that of the theory from which they were derived.

These results also seem to demonstrate the usefulness of orthogonal contrasts among categories as a way of combining categories to yield composites which are more reliable and interpretable than raw category frequencies, without losing whatever experimental independence the categories may have had. The use of factor analysis in conjunction with such contrasts provides a simple way

of generating "better" contrasts (i.e., ones which provide "better" descriptions of classroom behavior) than the original ones.

Analysis of Variance of Factor Scores

The 280 scores of the 70 teachers on each of the 15 factor keys were submitted to an analysis of variance in the same design that was used in estimating reliabilities of the OScAR contrasts and FIAT measures (see Table 15). In this application, several hypotheses were tested about the behaviors of the teachers, some related to the stability and some to the stability of the factor scores. Tests of these hypotheses were based on assumptions that Change and Subject were fixed and that Visit and Teacher were random in the population. Results of tests are shown in Table 27.

Insert Table 27 about here

Changes in Behavior.

The hypothesis that there was no stable change in teacher behavior between February and May was accepted with respect to eight behavior dimensions and rejected with respect to seven, as indicated in the column headed "Change." Rejection of this hypothesis required, in each instance, that variation between visits one semester apart in time be significantly larger than variation between visits one week apart, and justified the conclusion that a stable change in the factor key in question took place between February and May.

Figure 3 shows these changes in graphic form. To facilitate

Table 27

Summary of Results of Analyses of Variance of

Scores on 15 Factor-Derived Behavior Dimensions

Factor	Variable	Instrument	Hypotheses Rejected						Reliability Estimates	
			Change	Subject	Change by Subject	Teacher	Change by Teacher	Visit by Teacher	Total Scores	Change Scores
I	Lecturing	OScAR	*			**			.83	--
I	Lecturing	FIAT		*		**	**		.79	.42
II	Question Type	OScAR				**	**		.77	.48
II	Modified I-D Contrast	FIAT				**	**		.86	.40
III	Question Difficulty	OScAR	**			**			.49	--
IV	Student Initiations	FIAT	**			**	*		.80	.39
V	Rebuking	OScAR				**			.83	--
V	Criticizing	FIAT				**	**	*	.82	.40
VI	Listening	OScAR		*		**	**		.80	.82
VI	Modified Content Cross	FIAT	*			**	**		.86	.58
VII	Indirect Teacher Talk	FIAT		*		**	**		.71	.45
III	Question Source	OScAR				**			.79	--
VIII	Student Interruptions	FIAT	**			**	*		.67	.31
IX	Permissive Behavior	OScAR	**			**			.50	--
X	Managing Behavior	OScAR	*			**	**		.45	.33

*$.01 < P < .05$

**$P < .01$

Insert Figure 3 about here

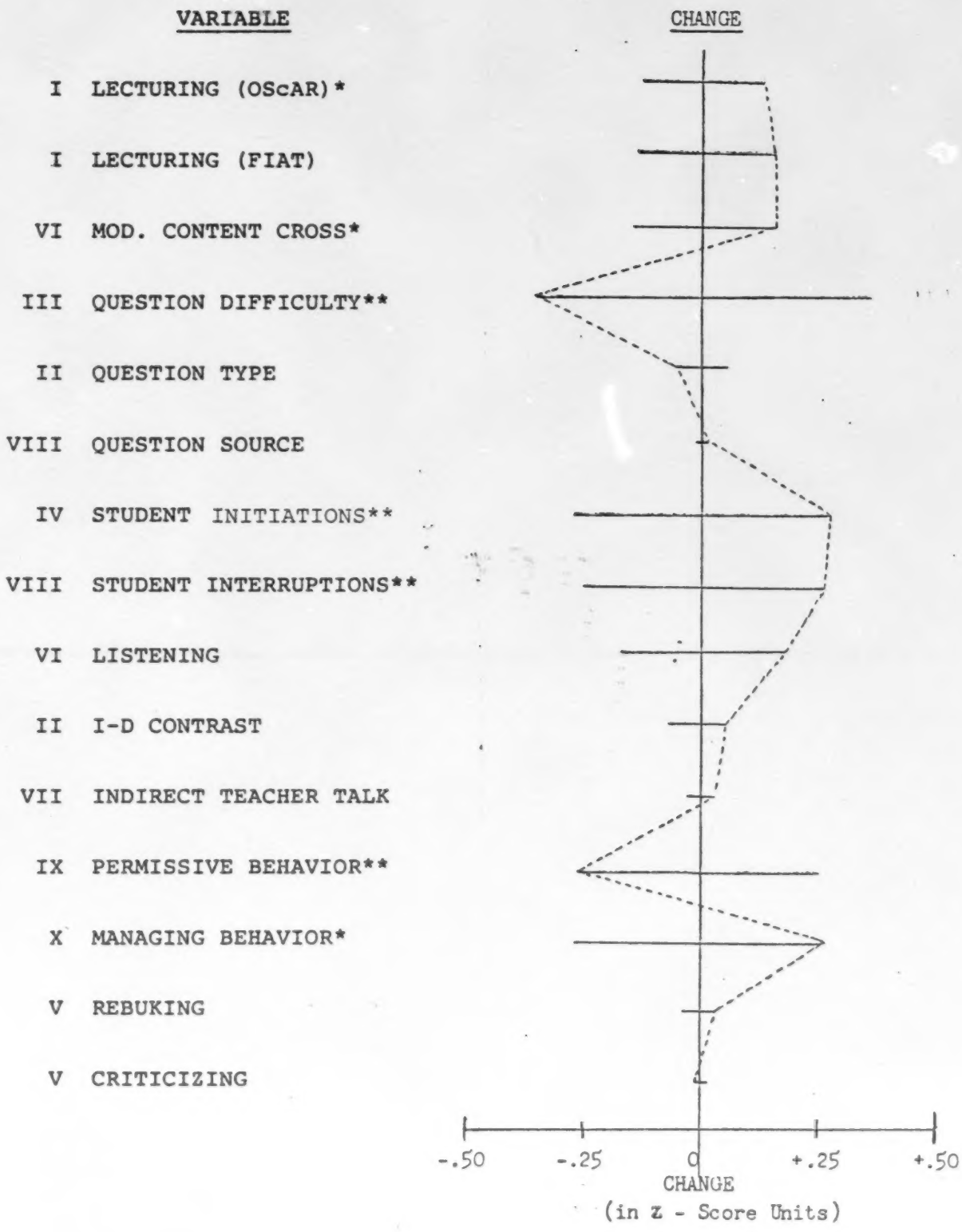
interpretation, each change was "standardized" by dividing its deviation from the sample mean by the standard deviation of the distribution of scores of the entire sample of teachers. The order of the dimensions (whether significant or not) was rearranged as shown. (The dotted line in the figure represents the "post" profile.)

The first three dimensions, which all measure much the same thing--lecturing vs. interaction with pupils--increased uniformly by about .3 of a standard deviation unit. This suggests that, as the year wore on, the teachers tended to talk more and to interact with pupils less.

The next three dimensions plotted relate to questioning behavior. It seems that the questions the teachers asked tended to become more difficult as time went on--or at least, that students tended to give fewer correct answers to them. The type of question asked and the relative number of questions raised by students showed no appreciable change, however.

Student Initiations (Category 9 on FIAT) and Student Interruptions (FIAT Cell 59) both increased sharply in number. There also appeared to be an upward trend in the teachers' willingness to listen to students without interrupting them, although the increase was not great enough to rule out change as an explanation.

Both measures of indirect behavior used in this study proved



* $.01 < P < .05$
** $P < .01$

Figure 3. Changes in teachers' scores on 15 factor keys

resistant to change, as did the only two affect-related measures used (Rebuking and Criticizing). These characteristics of teacher behavior seem to have rather deep roots and do not change easily--at least not without some sort of intervention.

The decline in Permissive Behavior and the increase in Managing Behavior suggest that even during their second semester in the classroom beginning teachers may still be tightening up their control of students--or trying to do so, at least.

In interpreting the findings shown in Figure 3, it is important to consider the results obtained when the hypothesis of no interaction between teacher and change was tested for each key (cf. Table 27). The hypothesis was rejected for 10 of the 15 dimensions.

Rejection of this hypothesis implies that, with respect to the dimension in question, changes in different teachers of the same subject varied, either in direction or amount. On those variables which showed no significant main effect for Change, this implies that some teachers showed significant changes, but not all of them.

When no significant interaction is found but the main effect is significant, the indication is that all teachers changed uniformly on that dimension. This is true of three of the changes noted: apparently all of the teachers lectured more, asked more difficult questions, and were less permissive at the end of the semester than at its beginning. These changes are the ones which are most likely to be the result of experience; they may be sum-

marized as reflecting an attempt to increase teacher initiative and control.

Acceptance of both hypotheses implies that none of the teachers changed appreciably on the dimension in question. This was true of Rebuking Behavior and Question Source, which seem to be the two most stable characteristics of these teachers as a group.

The affective dimensions on FIAT (I-D Contrast, Indirect Teacher Talk, and Criticizing) all fall into a category in which there was no uniform change, but in which at least some teachers did change. Because the mean changes are themselves so small, it appears that some teachers moved in one direction on these dimensions, others in the opposite one. These are behaviors, then, with respect to which changes are idiosyncratic. In trying to achieve their common purpose of self-improvement by trial and error, some teachers seem to choose opposite strategies.

Remaining are a group of dimensions on which both the main effect and interaction were significant. These are dimensions on which all teachers changed, but some more than others. Notable among these are Student Interruptions, Student Initiations, and Managing Behavior.

Thus to describe what happens to teachers during their second semester of full time teaching experience solely in terms of Figure 3 is to oversimplify things considerably. Teachers react for the most part as individuals, some changing in one way, some in the opposite one, some not at all.

Subject-Matter Differences

Significant variation between behaviors of teachers of different subjects was observed on three of the 15 scales (cf. Table 27). Significance in this instance means that the variation between teachers of different subjects (recorded by the same observer) exceeded variation between teachers of the same subject (recorded by the same observer). In the analysis of variance of scores on each factor key (see Table 15) the six degrees of freedom associated with between-subjects variation, which reflect six contrasts between subjects, one per observer (cf. Table 12), were pooled to yield a "between subjects within observers" mean square (line 4, Table 5). This was something of a hodge-podge, since the two subject areas contrasted varied from observer to observer; but it was nevertheless tested in each analysis (against variance between teachers within subjects and observers, in line 8). Only when this combined mean square was statistically significant was the null hypothesis rejected. In each of the three analyses in which the null hypothesis was rejected, the mean square associated with each of the six degrees of freedom was calculated and tested separately to find out which of the six contrasts were significant. With the three contrasts used (which grew out of the distribution of subjects in the class of interns studied), it is also possible to infer the rank order of the groups in any behavior dimension because science and English teachers appear in two contrasts each--once in the same and once in different ones. All of the pair-wise differences are shown in Table 28.

Insert Table 28 about here

The most clear-cut and consistent differences between subject-matter groups appeared on Lecturing Behavior (as scored on FIAT); these results are shown in the first column of Table 28. English teachers in this sample lectured less--and interacted with pupils more--than science teachers at both the junior and senior high school levels. At the junior high school level, social studies teachers also interacted more with students, and lectured less, than science teachers. Only the senior high school social studies teachers could compare with the science teachers when it came to lecturing!

Senior high school English teachers exhibited significantly more Listening Behavior than senior high mathematics teachers. None of the senior high teachers of other subjects than mathematics differed much from the English teachers in this regard, nor were any differences observed among teachers of different subjects at the junior high school level.

With respect to Indirect Teacher Talk, science teachers at the senior high level were significantly higher than social studies teachers at the same level, and more indirect than any other group, in fact.

It should be noted that no interaction was found between change and subject on any dimension; if teachers of different subjects tend to react to experience in different ways, the tendency

Table 28

Differences Between Mean Scores of Teachers of Different Pairs of Subjects at Different Levels on Three Behavior Dimensions

Level	Subjects	Dimension		
		Lecturing	Listening	Indirect Teacher Talk
Junior High	English - Mathematics	-35	+0.2	+0.6
	English - Science	-56*	+2.5	+9.1
	Social Studies - Science	-73*	+0.4	+1.1
Senior High	English - Mathematics	-3	+22.7*	-0.5
	English - Science	-56*	+3,7	-7.8
	Social Studies - Science	+8	-0.3	-15.0*

*$p < .05$

was not reflected in these data.

Reliability of Change Scores

The statistics shown in the last column in Table 28 are intraclass correlations estimating the reliabilities of change scores based upon four visits. They are reported only for those factor scales which detected individual differences in behavior changes over the semester. Except in the case of Listening Behavior, these are quite low, reflecting the general experience that behavior change is much more difficult to measure reliably than behavior status. However, a significant reliability coefficient in this column does warrant the conclusion that at least some teachers changed their behavior on the dimension in question--or, to put it differently, that the rank order of the teachers changed on the dimension during the semester.

There were two dimensions on which (a) the average teacher did not change appreciably, and (b) there was no detectable difference in the amount of change in different teachers. In short, there was no evidence that any teacher changed at all. These, as has been suggested earlier, seem to be the most stable behaviors of all. These two dimensions were Rebuking Behaviors and Question Source. It should be remembered that Rebuking Behaviors here refers to reprimands that do not seem to reflect hostility or anger but rather to have a controlling function, and that Question Source measures (a) the relative number of questions asked (and unsolicited comments made) by pupils compared to the number asked by the teacher, and (b) any tendency of the teacher to respond differently

to pupils' voluntary comments than to those made in response to teacher questions.

The three dimensions on which all teachers changed the same amount seem to be ones with respect to which one could predict that any teacher would change with experience. All teachers in the population represented here tended to talk more, ask more difficult questions, and act less permissive at the end of their first year than in the middle of the year.

The four dimensions which showed significance on both tests are ones with respect to which there seems to be a strong pressure to change, but one which some teachers are more able to resist than others. What may well be a general tendency for pupils to volunteer more information and to interrupt the teacher more often as the school year nears its end seems to be resisted or quelled more readily by some teachers than by others. Also, although the tendency is for all teachers to devote more attention to management as time goes on, this tendency is greater for some than for others.

The largest single group of dimensions are those with respect to which there were significant changes, but not all in the same direction. Indirect Teacher Talk, Criticism, Question Type, Listening, and Lecturing (as measured on FIAT) fall in this category.

The general picture is one of teachers still engaged in experimentation in search of their own individual teaching styles. Insofar as any uniform change can be recognized, one can hardly

describe it as desirable. Perhaps what these teachers need is feedback about what is happening to them. Information of the kind presented here could be made available to the teachers themselves during this difficult first year in the classroom, possibly accompanied by suggestions as to what they might do to help themselves change in ways they regard as constructive. It is doubtful indeed whether many of these teachers either realize how their behavior is changing or would regard the changes as desirable if they did.

Cognitive Factors in Teacher Behavior

Fifty-three of the 70 teachers in the study had taken a form of the National Teacher Examinations, Common Examinations (NTE), before being admitted to the program, and it was possible to retrieve their answer sheets from storage at Educational Testing Service. This section will describe a study of the relationships between the performance of these 53 teachers on the examinations and their teaching style as measured by total scores on the OScAR and FIAT factor keys.

The Examination

The form of the Common Examinations of the NTE taken by the 53 teachers contained 345 multiple-choice items. The 345 items had been written according to a table of specifications which called for items representing 19 different content areas. About half of the items were designed to sample the teacher's knowledge of subject-matter commonly included in school curricula, such as science, mathematics, English, etc. The other half were designed to measure knowledge of the content of professional education courses--

history and philosophy of education, teaching principles and practices, and so on. Subscores were obtained for each teacher on each of 19 "subtests" made up of items from one of the 19 content areas.

Analysis of Internal Structure of the Examination

Answer sheets were available for 91 teachers, including the 53 who were observed plus 38 others enrolled in the same program but not observed in their classrooms. These 91 papers were submitted to an analysis of variance of the form originally suggested by Hoyt (1941), to study the internal structure of the test. Hoyt used a two-way design without replication, items by candidates; in this case, there were 90 degrees of freedom for candidates, 344 for items, and 30,960 for error. Hoyt's design was extended by partitioning the 344 degrees of freedom for items into two portions as shown in Table 29. One portion, with 18 degrees of free-

---

Insert Table 29 about here

---

dom (Line 2), estimated between-items variance from variation between items on different subtests only, the other, with 326 degrees of freedom (Line 3), estimated the same type of variance by comparing only items on the same subtest.

As shown in Lines 4 and 5 of Table 29, the sum of squares for error (which would have had 30,960 degrees of freedom in a conventional Hoyt analysis) was partitioned in an analogous fashion. One portion, with 1,620 degrees of freedom, estimated errors of

Table 29

Analysis of the Performance of the
91 Teachers on 19 Subtests of the
Common Examinations of the National Teacher Examinations

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square Obtained	Mean Square Expected
1. Candidates	90	253.20	2.81	$345\sigma^2_c + \sigma^2$
2. Subtests	18	348.64	19.37	$91\bar{K}\theta^2_s + 91\sigma^2_i + \sigma^2$
3. Items (in subtests)	326	1,202.18	3.69	$91\sigma^2_i + \sigma^2$
4. Interaction, Candidate by Subtest	1,620	528.49	0.33	$\bar{K}\sigma^2_{cs} + \sigma^2$
5. Residual	29,340	5,089.91	0.17	σ^2
6. Total Variation	31,394	7,422.41		

$$\bar{K} = \frac{1}{18}\left(345 - \frac{\Sigma K^2_i}{345}\right)$$

K_i = number of items on subtest i

i = 1, 2, ..., 19

measurement from interaction between candidates and items on different subtests, and the other, with 29,340 degrees of freedom, estimated errors of measurement from interaction between candidates and items on the same subtest.

The fact that the former mean square is larger than the latter indicates that there is an interaction between candidates and subtests. The null hypothesis that all of the 19 subtests are measuring the same function may therefore be rejected; if some candidates tend to do better on one subtest than they do on others, the subtests cannot be measuring the same knowledge or ability. If they were measuring the same function--if they were equivalent--then the order of the true scores of the 91 candidates would not vary from one subtest to another. The practical conclusion is that there is information in the subtest scores that does not appear in the total score, so that at least some of the subscores should be retained for further analysis.

The question remains: how many and which subtest scores should be retained? To answer this question the 18 degrees of freedom between subtests and the 1,620 for candidate by subtest interaction were partitioned into 18 parts each. In the case of subtests there was one degree of freedom for each portion; in the case of interaction, there were 90 degrees of freedom for each portion. Each portion was based on one orthogonal contrast between two sets of subtests, each reflecting one of 18 a priori hypotheses implicit in the test specification about how the content areas sampled by the 19 subtests might differ.

The null hypothesis was rejected in eight instances and was accepted in ten, as regards between-subtests variation. The null hypothesis was rejected in nine instances and accepted in nine as regards interaction. Table 30 presents a version of the complete

---

Insert Table 30 about here

---

analysis of variance in which all non-significant mean squares have been pooled with their respective error terms.

In order to conserve all of the information in the test scores it was therefore necessary to retain scores on the 11 subtests in Table 31. Table 31 also shows the number of items and the mean reliability per item of each subtest.

Insert Table 31 about here

This last statistic is equivalent to what would be obtained by using the Spearman-Brown formula (Gulliksen, 1950, pp. 77-79) backwards on each subtest, "prophesying" the reliability of a one-item test in each instance (cf. Cronbach, 1951). The mean reliability per item gives a pretty good idea of the extent to which each subtest is saturated with its own principal component, since its magnitude is independent of the number of items on the subtest.

Regression on Behavior Dimensions

Each of the 15 behavior dimensions scored on OScAR and FIAT

Table 30

Pooled Analysis of Variance
of Scores of 91 Teachers on 345 Items of the
Common Examinations of the National Teacher Examinations

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Candidates	90	253.20	2.81
General Knowledge (vs. Professional)	1	96.30	96.30
Science and Mathematics (vs. rest of General Knowledge)	1	17.07	17.07
Science (vs. Mathematics)	1	18.46	18.46
English (vs. Social Studies and Fine Arts)	1	45.91	45.91
Literature (vs. English)	1	35.46	35.46
Social Studies (vs. Fine Arts)	1	16.32	16.32
Foundations (vs. Teaching Principles and Practices)	1	27.61	27.61
History and Philosophy (vs. School and Society)	1	48.84	48.84
Items (within Subsets)	336	1,244.85	3.71
Candidate x General (vs. Professional Knowledge)	90	57.29	0.64
Candidate x Science and Mathematics (vs. Other General Knowledge)	90	78.17	0.87
Candidate x Science (vs. Mathematics)	90	42.62	0.47
Candidate x English (vs. Social Studies and Fine Arts)	90	41.15	0.46
Candidate x Literature (vs. English Mechanics and Effectiveness)	90	53.44	0.59
Candidate x English A (vs. English B)	90	23.00	0.26
Candidate x Social Studies (vs. Fine Arts)	90	44.42	0.49
Candidate x Foundations (vs. Teaching Principles and Practices)	90	20.39	0.23
Candidate x History and Philosophy (vs. School and Society)	90	22.18	0.25
Pooled Residual	30,150	5,235.73	0.17
Totals	31,394	7,422.41	

Table 31

Subtests of NTE Common Examinations

Retained for Analysis

Subtest	Number of Items	Mean Reliability Per Item
Science	30	.16
Mathematics	20	.20
English A	30	.11
English B	25	.09
Literature	19	.21
Social Studies	30	.08
Fine Arts	16	.18
Teaching Principles and Practices	60	.03
History and Philosophy of Education	20	.03
School and Society	22	.02
Psychological Foundations plus Teacher Role plus School Organization	73	.04
Total	345	

in turn was regressed on all 11 NTE subtests. Eight of the equations obtained are shown in Table 32. Neither the multiple

---

Insert Table 32 about here

---

correlations nor any of the beta weights in any of the other seven equations was significantly different from zero, so they are not shown.

Since only two of the 15 equations resulted in multiple correlations whose probability under the null hypothesis was less than .05, and since only nine of the 165 beta weights met this criterion, these findings should be regarded as tentative only. Because data of this type are so rare, however, they may be worth a closer look.

Attempts to predict teacher competence (as measured by various criteria) from teachers' scores on cognitive tests have generally been unsuccessful in the past (cf. Barr, 1948). Here a different question has been raised: instead of asking whether teachers' test scores correlate with some ill-defined and poorly measured construct called teacher effectiveness, this study asks whether they correlate with stable patterns of classroom behavior which may be regarded as elements of teacher style, and which are clearly defined in operational terms.

The one impressive finding in Table 32 is the multiple correlation of .66 between Lecturing Behavior (as scored on the Flanders' records) and performance on the NTE. Inspection of the beta

Table 32

Regressions of Teacher Behaviors
on National Teacher Examinations Common Examinations

Dependent Variables	Multiple R	Beta Weights on Test Subscores										
		Science	Mathematics	English A	English B	Literature	Social Studies	Fine Arts	Teaching Principles	History and Philosophy	School and Society	Psychology etc.
1. Lecturing Behavior (FIAT)	.66*	+.38*	+.13	+.13	+.18	-.37	-.15	+.16	-.42*	+.30	+.02	-.04
2. Lecturing Behavior (OScAR)	.56	+.14	+.05	+.19	+.19	-.30	+.30	+.20	-.42*	-.05	+.15	-.25
3. Modified Content Cross (FIAT)	.60*	+.39*	-.21	+.31	+.05	-.17	-.01	+.17	-.12	+.25	-.14	+.02
4. Criticizing Behavior (FIAT)	.55	-.02	-.12	-.48*	+.19	-.30	-.09	-.16	-.09	+.33	+.38	+.11
5. Rebuking Behavior (OScAR)	.48	-.06	-.18	-.27	+.26	-.52*	-.07	+.07	+.01	+.47	+.33	+.02
6. I-D Contrast on Student Response (FIAT)	.53	+.36	-.28	+.16	+.19	-.11	-.10	+.08	-.03	+.42*	+.00	-.18
7. Question Quality (OScAR)	.53	-.04	+.27	-.19	-.33	+.20	-.09	-.08	-.27	-.09	+.46*	-.12
8. Listening Behavior (OScAR)	.48	-.19	+.25	-.05	-.46*	+.36	+.11	+.07	+.21	-.09	-.26	+.10

*P < .05

weights in the equation indicates that the scores a teacher obtains on (a) science items, and (b) items related to teaching principles and practices, are principally responsible for this relationship. Teachers who do better on the science items lecture more; teachers who do better on the teaching principles and practices items lecture less.

Results of the analysis of factor scores indicate that science teachers as a group tend to lecture more than other teachers (cf. Table 28 above). This suggests that the contribution of the science subtest to the regression equation may be a function of subject taught, in part at least. However, since there were only seven science teachers among the 53 included in this analysis, it is likely that teachers of other subjects who had high science subtest scores also tended to act like science teachers no matter what subject they taught.

The negative relationship between lecturing and knowledge of items related to teaching principles and practices is intriguing, suggesting as it does that the teacher who lectures may do so only because he does not know any other way to teach!

Discussion

The rest of the results in Table 32 should be looked upon as only suggestive. In doing this it may be useful to ask the question: if a teacher gets his highest score on items of one particular type, what kind of teaching behavior would he be likely to exhibit in his classroom?

The teacher scoring highest on science items would tend to

be high on Lecturing Behavior, as has been noted, and also on the Modified Content Cross and the I-D Contrast on Student Responses.

What these dimensions all have in common is a sensitivity to two of Flanders' ten categories: Lecturing and Asking Questions, plus a negative weighting on pupil responses, particularly those to which the teacher reacts in a direct fashion. The teacher whose forte is science spends a lot of time dealing with subject-matter, and tends to dominate the discussion himself, though perhaps in an accepting (if not supportive) manner.

No patterns emerge for teachers with high scores on mathematics or English A. Those who score highest on English B, which was designed to measure effectiveness of expression, are low on Listening. This means that pupil comments in their classroom tend to be brief; their students do not deliver monologues or speak at length without teacher interruption.

The teacher who scores highest on literature items tends to behave in the opposite fashion. He listens to his pupils more, and lets them speak at length; he seldom finds cause to rebuke them, and his own comments tend to be brief.

Teachers scoring high on social studies and art items show no clear-cut pattern; as far as subject-matter content goes, then, it appears to be scores on the science, effectiveness of expression, and literary acquaintance items that relate to teaching style.

Inspection of the subtests related to professional knowledge suggests that teachers who know most about teaching principles and

practices may be expected to teach by question-and-answer rather than by the lecture method. (This is confirmed on both of the Lecturing measures.)

The teacher who scores highest on the 20 items devoted to the history and philosophy of education also scores high on Rebuking and on the I-D contrast on student response; the one who scores highest on items related to school and society tends to ask interrelated questions and to have students evaluate them, rather than asking short "drill-type" questions.

Knowledge of psychology does not seem to be related to teaching style.

The tentative nature of these inferences prevents their being taken seriously as substantive findings; however, they do illustrate the kind of findings this line of investigation can produce.

Summary and Conclusions

Further advances in the scientific study of teaching--as distinguished from tutoring and instructing--depend on the development of practicable, objective procedures for measuring teacher behavior, for describing the teaching process in quantitative language to facilitate the use of the scientific method. The project reported above was designed to test the potential usefulness for this purpose of the OScAR 4V-- Observation Schedule and Record, Form 4 Verbal.

Four visits were made to the classrooms of each of 70 secondary level intern teachers, two visits being made during February and two during May of the same year, by teams of two trained observers. One observer recorded interactions by the familiar Flander's Interaction Analysis Techniques (FIAT); the other recorded the same interactions using OScAR 4V.

Each OScAR record consisted of a vector of 42 frequencies, one for each of 42 categories of events recognizable in the record. These were transformed into 41 experimentally independent a priori linear contrasts. Each FIAT record was scored on 44 "measures" which consisted of various linear combinations of cells from th I. matrix based on various "ratios" reported in the FIAT literature.

Stability coefficients on the 41 OScAR contrasts ranged from zero to .85 with a median of .59; stability coefficient on the 44 FIAT measures ranged from .07 to .85 with a median of .68. The 85 variables were intercorrelated across the 280 visits, the first ten principal components were extracted, and the loadings were rotated to simple structure.

Eight factors showed high loadings on OScAR contrasts, of which five also showed high loadings on FIAT measures. Two factors--Student Initiations and Indirect Teacher Talk--had high loadings on FIAT measures but not on OScAR contrasts.

Simplified factor keys were developed for OScAR by adding weights of those contrasts which had high loadings on each factor together (with plus or minus signs according to the sign of the loading). Since there was no overlap between keys, this yielded eight

experimentally independent (i.e., non-overlapping) keys or dimensions of behavior to be derived from an OScAR record.

Three of the dimensions--Rebuking, Managing, and Permissive Behavior--seem to reflect primarily how the teacher controls the sequence of events in the classroom. The other five appear to reflect the cognitive climate--Lecturing, Question Difficulty, Question Quality, Question Source, and Listening. The stability of these eight keys ranged from .45 (Managing) to .83 (Lecturing and Rebuking) with a median of .78, when based on four visits of about 28 minutes each by one observer with six hours of training.

An analysis of variance of the scores of the 70 teachers revealed that the dimensions were sensitive to uniform changes in teaching style over time, to stable changes in the behavior of individual teachers, and (in the case of the Listening key only) to differences between teachers of different subjects.

Fifty-three of the 70 teachers had taken the Common Examination of the National Teacher Examination so an analysis of relationships between teaching style and performance on the examination was carried out. There was some tendency for knowledge of certain areas of content to relate to some of the dimensions of behavior measured on OScAR, but not enough to establish any relationship beyond chance.

References

Allen, Graham. Teaching behaviour and pupils' number development. Hawthorn, Victoria: Australian Council for Educational Research, 1973.

Amidon, E. J., & Flanders, N. The role of the teacher in the classroom. Minneapolis, Minnesota: Paul S. Amidon Associates, Inc., 1963. In E. J. Amidon & J. B. Hough (eds.), Interaction Analysis: Theory, Research, and Application. Reading, Massachusetts: Addison-Wesley, 1967.

Amidon, E. J., & Hough, J. B. Interaction analysis: Theory, reseach, and application. Reading, Massachusetts: Addison-Wesley, 1967.

Amidon, E. J., & Hunter, Elizabeth. Improving teaching: Analyzing verbal interaction in the classroom. New York: Holt, Rhinehart and Winston, 1966.

Barr, A. S. The measurement and prediction of teaching efficiency: A summary of investigation. Journal of Experimental Education, 1948, 16, 203-283.

Bowers, N. D., & Soar, R. S. Evaluating human relations training for classroom teachers. Educational Leadership, 1960, 17, 451-454.

Bowers, N. D., & Soar, R. S. Studies of human relations in the teaching-learning process. Final Report: Evaluative of laboratory human relations training for classroom teachers. Chapel Hill, North Carolina (United States Office of Education Cooperative Research Project No. 469), 1961.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Flanders, N. A. Interaction analysis in the classroom: A manual for observers. Minneapolis; College of Education, University of Minnesota, 1960.

Flanders, N. A. Teacher influence, pupil attitudes, and achievement. Cooperative Research Monograph, No. 12. Washington, D. C.: United States Government Printing Office, 1965.

Flanders, N. A. Analyzing teacher behavior. Reading, Massachusetts: Addison-Wesley, 1970.

Fowler, Beverly. Relation of teacher personality characteristics and attitudes to teacher-pupil rapport and emotional climate in the elementary classroom. Doctoral dissertation, University of South Carolina, University Microfilms, 1962.

Fowler, Beverly. Use of OScAR in a study to determine the relationship between emotional climate and teacher personality characteristics and attitudes. Paper presented at the National Society for College Teachers, Educational Psychology Section, February 17, 1966.

Gage, N. L. (Ed.) Handbook of research on teaching. The American Educational Research Association, Rand McNally and Company, Chicago, 1963.

Green, B. F., Jr., & Tukey, J. W. Complex analysis of variance: General problems. Psychometrika, 1960, 25, 127-152.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hotelling, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 1933, 24, 417-441, 498-520.

Hoyt, C. J. Test reliability estimated by analyses of variance. Psychometrika, 1941, 6, 153-160.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. Psychometrika, 1958, 23, 187-200.

Li, C. C. Introduction to experimental statistics. New York: McGraw-Hill, 1964.

Medley, D. M. Cognitive factors in teaching style. Paper presented at the meeting of the American Educational Research Association, Minneapolis, Minnesota, 1970.

Medley, D. M. Language of teacher behavior. Journal of Teacher Education, 1971, 22, 157-165.

Medley, D. M., & Hill, R. A. A comparison of two techniques for analyzing classroom behaviors. American Educational Research Assoication, Chicago, Illinois, 1968. In R. C. Burkhart (Ed.), The Assessment Revolution: New Viewpoints for Teacher Evaluation, Albany, New York State Education Department, Division of Teacher Education and Certification, 1969.

Medley, D. M., & Impelleteri, J. T., & Smith, Lou. Coding teacher's verbal behavior in the classroom: A manual for users of OScAR 4V. New York: Division of Teacher Education of the City University of New York, 1966.

Medley, D. M., & Mitzel, H. E. Studies of teacher behavior: Refinement of two techniques for assessing teacher's classroom behaviors. New York: Board of Higher Education, City of New York, Division of Teacher Education, 1955. (Research Series No. 28).

Medley, D. M., & Mitzel, H. E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1958, 49, 86-92.

Medley, D. M., & Mitzel, H. E. Some behavioral correlates of teacher effectiveness. Journal of Educational Psychology, 1959, 50, 239-246.

Medely, D. M., & Mitzel, H. E. Measured changes in student teaching behavior, Chapter 4 in Schueler, H., Gold, M., & Mitzel, H. E. The Uses of Television for Improving Measures of Student Teaching Performance. New York: Hunter College of City University of New York (Phase I, United States Office of Education Grant No. 730035), 1962.

Medley, D. M., & Mitzel, H. E. A tentative framework for the study of effective teacher behavior. Journal of Experimental Education, 1962, 30, (4), 318-320.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observations. In N. L. Gage (Ed.), Handbook of Research on Teaching, Chicago, Illinois: Rand McNally, 1963, pp. 247-328.

Medley, D. M., & Mitzel, H. E. Development of the observation schedule. Chapter 3 in Schueler, H., Gold, M., & Mitzel, H. E. The Use of Television for Improving Measures of Student Teaching Performance. New York: Hunter College of the City University of New York (Phase I, United States Office of Education Grant No. 730035), 1962.

Mitzel, H. E., & Rabinowitz, W. Assessing social-emotional climate in the classroom by Withall's techniques. Psychological Monographs, 1953, 67, No. 18, (Whole No. 368).

Morrison, Virginia B. The relationship of student teacher performance and pupil performance to supervisory and pupil merit ratings. Unpublished doctroal dissertation, University of Michigan, 1961.

Morrison, Virginia B. Research in the teaching process: Video-taped observations using the OScAR technique. Paper presented at the National Society for College Teachers, February, 1966.

Morrison, Virginia B. Teacher-pupil interaction in three types of classroom reading situations. Detroit, Michigan: Wayne State University, 1967.

Schueler, H., Gold, M., & Mitzel, H. E. The use of television for improving measures of student teaching performance. New York: Hunter College of the City University of New York (Phase I, United States Office of Education Grant No. 730035), 1962.

Simon, Anita & Boyer G. E. (Eds.) Mirrors for behavior: An anthology of classroom observation instruments. Philadelphia Research for Better Schools, 1967.

Soar, R. S. Multivariate statistical procedures in predicting teacher-pupil classroom behavior. Columbia, South Carolina (United States Office of Educational Cooperative Research Project No. 1170), 1962.

Winer, E. J. Statistical principals in experimental design. New York: McGraw-Hill, 1962.

Withall, J. G. The development of a technique for the measurement of social-emotional climate in classrooms. Unpublished doctoral dissertation, Department of Education, University of Chicago, 1948.

Withall, J. G. Development of a technique for the measurement of socio-emotional climate in classrooms. Journal of Experimental Education, 1949, 17, 347-361.

Withall, J. G. The development of a climate index. Journal of Educational Psychology, 1951, 45, 93-99.

Withall, J. G. An objective measurement of a teacher's classroom interactions. Journal of Educational Psychology, 1956, 47, 203-212.

OBSERVATION SCHEDULE AND RECORD (OScAR 4V)

55

DONALD M. MEDLEY

Educational Testing Service
Princeton, New Jersey

This category system has affective, cognitive and procedural dimensions which reveals how much of the time the teacher and pupils are spending on areas other than the content matter of the classroom. It has dimensions for analyzing both teacher and pupil verbal behaviors. The system codes two sets of behaviors: interchanges and monologues.

The interchanges report only the teacher's behavior, making note of how the teacher starts an interchange and then how he responds to the pupil's answer. This technique retains some of the sequential nature of the data, and in fact has the coder build something similar to a Flanders Matrix as he codes. By scanning the finished coding sheet, the reader can tell immediately whether the teacher responded with explicit or implicit praise or rejection, responded neutrally, or ignored four kinds of pupil comments:

1) pupil questions,
2) responses to questions which refer directly to a previous pupil's comment,
3) responses to questions which offer a pupil a choice of more than one acceptable answer or,
4) responses to questions which have only one acceptable answer.

This system, along with the Flanders System, has been used in an inter-observation system comparison to determine the differences resulting between Flanders and OScAR data. Cross system comparison is an important step for providing data necessary to create an optimal set of categories for use in varying settings. The author recommends an OScAR observation period ranging from 30 minutes to a full class period.

OBSERVATION SCHEDULE AND RECORD (OScAR 4V)

55

SUBJECT OF OBSERVATION

- Teacher and Pupils
  Teacher Only
  Pupil Only
  Small Groups
  Family Dyads
  Counselor or Therapist with Patient
  Administrators/Supervisors and Supervisees

SETTINGS IN WHICH USED

- Classroom, any content
  Classroom, for specific subject
  Commercial or Industrial
  Counseling or Therapy
  Group Dynamics
  Other

NUMBER OF SUBJECTS OBSERVED

1 Only
Dyad
More Than 2 People But Not Classroom Setting
- More Than 2 People in Classroom Setting
  Point Time Sample

CODING UNITS

- Category Change
  Time Unit
  Topic or Content Change
- Speaker Change
  Time Sample
- Other *question-answer-response unit*

COLLECTION METHODS REPORTED

- Live *(no special equipment needed)*
  Live *(special coding equipment needed)*
  Video and/or Audio Tape Required

COLLECTING AND CODING PERSONNEL NEEDED

- One Person Only
  Team of Two
  2 Teams of Two

CATEGORY DIMENSIONS OF THE SYSTEM

- Affective
- Cognitive
- Procedure or Routine
  Physical Environment *(material, equipment, etc.)*
  Psychomotor *(body movement)*
  Activity *(doing something)*
  Sociological Structure *(role, who to whom, etc.)*
  Other

USES REPORTED BY AUTHOR

- Research
  Training
  Evaluation

CATEGORIES FOR

VERBAL BEHAVIOR IN OScAR 4V

D. M. Medley, J. T. Impellitteri, L. H. Smith

I. STATEMENTS

A. Teacher statements--utterances which neither respond to nor solicit a response from a pupil--are classified as follows:

1. AFFECTIVE. A statement revealing sensitivity to pupil feelings is classified as CONSIDERING. A statement criticizing pupil conduct is classified as REBUKING.

2. SUBSTANTIVE. A statement containing no affect but referring directly to content to be learned by pupils is classified as INFORMING if it conveys a fact, generalization, or the like, or PROBLEM STRUCTURING if it sets up a question or issue to be solved.

3. PROCEDURAL. A statement which contains neither affect nor substance is classified as DIRECTIVE if it contains a command or an instruction with the force of a command. A statement which does not clearly fall into one of the above categories is classified as DESCRIBING.

B. Pupil statements--utterances by pupils addressed to other pupils are classified as PUPIL STATEMENTS.

C. Sequence. If a teacher makes two or more successive statements which may be classified in the same category, all except the first are classified as CONTINUING. The first statement in a series of the same kind is classified as INITIATING.

II. INTERCHANGES

An interchange is an episode in which a pupil says something to the teacher and the teacher reacts.

A. Substantive interchanges are those in which the pupil's utterance refers to content to be learned.

B. Entries. A substantive interchange begins with one of four types of entries:

1. PUPIL INITIATED. The pupil addresses a statement or question to the teacher.

2. ELABORATING. The teacher addresses a question to a pupil which refers directly to a previous pupil comment.

3. DIVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment, and which offers him a choice of two or more acceptable or "correct" answers.

4. CONVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment and to which there is only one acceptable answer.

C. Exits. After the pupil has asked his question or made his answer, the teacher disposes of the answer in one of six ways, called Exits. Exits are first classified according to the information they contain about the correctness or acceptability of what the pupil has said.

If the teacher clearly indicates that what the pupil has said is correct or acceptable, the interchange is classified as SUPPORTED if praise or enthusiasm is shown, as APPROVED if praise is not given.

If the teacher clearly indicates that what the pupil has said is incorrect or unacceptable, the interchange is classified as CRITICIZED if disapproval of either the pupil or what he has said is expressed, or as NEUTRALLY REJECTED if no disapproval is expressed.

If the teacher makes some response to what the pupil says which does not clearly indicate whether it is correct (acceptable) or incorrect (unacceptable) the interchange is classified as ACCEPTED; if the teacher makes no response, it is classified as NOT EVALUATED.

Since these four entries and six exits are used in classifying them, it may be said that 24 kinds of substantive interchanges are recognized.

D. Non-Substantive Interchanges are those in which the pupils contribution does not refer to content to be learned.

1. TEACHER-INITIATED non-substantive interchanges are classified as POSITIVE or NEGATIVE according to the affective content of the teacher's question.

2. PUPIL-INITIATED non-substantive interchanges are classified as POSITIVE if the teacher supports, approves, or accepts the pupil's suggestion, and as negative if he criticizes, neutrally rejects, or ignores it.

OBSERVATION SCHEDULE AND RECORD (OScAR 4V)

55

DONALD M. MEDLEY

Educational Testing Service
Princeton, New Jersey

This category system has affective, cognitive and procedural dimensions which reveals how much of the time the teacher and pupils are spending on areas other than the content matter of the classroom. It has dimensions for analyzing both teacher and pupil verbal behaviors. The system codes two sets of behaviors: interchanges and monologues.

The interchanges report only the teacher's behavior, making note of how the teacher starts an interchange and then how he responds to the pupil's answer. This technique retains some of the sequential nature of the data, and in fact has the coder build something similar to a Flanders Matrix as he codes. By scanning the finished coding sheet, the reader can tell immediately whether the teacher responded with explicit or implicit praise or rejection, responded neutrally, or ignored four kinds of pupil comments:

1) pupil questions,
2) responses to questions which refer directly to a previous pupil's comment,
3) responses to questions which offer a pupil a choice of more than one acceptable answer or,
4) responses to questions which have only one acceptable answer.

This system, along with the Flanders System, has been used in an inter-observation system comparison to determine the differences resulting between Flanders and OScAR data. Cross system comparison is an important step for providing data necessary to create an optimal set of categories for use in varying settings. The author recommends an OScAR observation period ranging from 30 minutes to a full class period.

OBSERVATION SCHEDULE AND RECORD (OScAR 4V)

55

SUBJECT OF OBSERVATION

- Teacher and Pupils
  Teacher Only
  Pupil Only
  Small Groups
  Family Dyads
  Counselor or Therapist with Patient
  Administrators/Supervisors and Supervisees

SETTINGS IN WHICH USED

- Classroom, any content
  Classroom, for specific subject
  Commercial or Industrial
  Counseling or Therapy
  Group Dynamics
  Other

NUMBER OF SUBJECTS OBSERVED

1 Only
Dyad
More Than 2 People But Not Classroom Setting
- More Than 2 People in Classroom Setting
  Point-Time Sample

CODING UNITS

- Category Change
  Time Unit
  Topic or Content Change
- Speaker Change
  Time Sample
- Other *question-answer-response unit*

COLLECTION METHODS REPORTED

- Live *(no special equipment needed)*
  Live *(special coding equipment needed)*
  Video and/or Audio Tape Required

COLLECTING AND CODING PERSONNEL NEEDED

- One Person Only
  Team of Two
  2 Teams of Two

CATEGORY DIMENSIONS OF THE SYSTEM

- Affective
- Cognitive
- Procedure or Routine
  Physical Environment *(material, equipment, etc.)*
  Psychomotor *(body movement)*
  Activity *(doing something)*
  Sociological Structure *(role, who to whom, etc.)*
  Other

USES REPORTED BY AUTHOR

- Research
  Training
  Evaluation

CATEGORIES FOR

VERBAL BEHAVIOR IN OScAR 4V

D. M. Medley, J. T. Impellitteri, L. H. Smith

I. STATEMENTS

A. Teacher statements--utterances which neither respond to nor solicit a response from a pupil--are classified as follows:

1. AFFECTIVE. A statement revealing sensitivity to pupil feelings is classified as CONSIDERING. A statement criticizing pupil conduct is classified as REBUKING.

2. SUBSTANTIVE. A statement containing no affect but referring directly to content to be learned by pupils is classified as INFORMING if it conveys a fact, generalization, or the like, or PROBLEM STRUCTURING if it sets up a question or issue to be solved.

3. PROCEDURAL. A statement which contains neither affect nor substance is classified as DIRECTIVE if it contains a command or an instruction with the force of a command. A statement which does not clearly fall into one of the above categories is classified as DESCRIBING.

B. Pupil statements--utterances by pupils addressed to other pupils are classified as PUPIL STATEMENTS.

C. Sequence. If a teacher makes two or more successive statements which may be classified in the same category, all except the first are classified as CONTINUING. The first statement in a series of the same kind is classified as INITIATING.

II. INTERCHANGES

An interchange is an episode in which a pupil says something to the teacher and the teacher reacts.

A. Substantive interchanges are those in which the pupil's utterance refers to content to be learned.

B. Entries. A substantive interchange begins with one of four types of entries:

1. PUPIL INITIATED. The pupil addresses a statement or question to the teacher.

2. ELABORATING. The teacher addresses a question to a pupil which refers directly to a previous pupil comment.

3. DIVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment, and which offers him a choice of two or more acceptable or "correct" answers.

4. CONVERGENT. The teacher addresses a question to a pupil which does not refer directly to a previous pupil comment and to which there is only one acceptable answer.

C. Exits. After the pupil has asked his question or made his answer, the teacher disposes of the answer in one of six ways, called Exits. Exits are first classified according to the information they contain about the correctness or acceptability of what the pupil has said.

If the teacher clearly indicates that what the pupil has said is correct or acceptable, the interchange is classified as SUPPORTED if praise or enthusiasm is shown, as APPROVED if praise is not given.

If the teacher clearly indicates that what the pupil has said is incorrect or unacceptable, the interchange is classified as CRITICIZED if disapproval of either the pupil or what he has said is expressed, or as NEUTRALLY REJECTED if no disapproval is expressed.

If the teacher makes some response to what the pupil says which does not clearly indicate whether it is correct (acceptable) or incorrect (unacceptable) the interchange is classified as ACCEPTED; if the teacher makes no response, it is classified as NOT EVALUATED.

Since these four entries and six exits are used in classifying them, it may be said that 24 kinds of substantive interchanges are recognized.

D. Non-Substantive Interchanges are those in which the pupils contribution does not refer to content to be learned.

1. TEACHER-INITIATED non-substantive interchanges are classified as POSITIVE or NEGATIVE according to the affective content of the teacher's question.

2. PUPIL-INITIATED non-substantive interchanges are classified as POSITIVE if the teacher supports, approves, or accepts the pupil's suggestion, and as negative if he criticizes, neutrally rejects, or ignores it.